Compressive Sampling of Ensembles of Correlated Signals

Ali Ahmed and Justin Romberg*

DRAFT: 1:24am, January 28, 2015

Abstract

We propose several sampling architectures for the efficient acquisition of an ensemble of correlated signals. We show that without prior knowledge of the correlation structure, each of our architectures (under different sets of assumptions) can acquire the ensemble at a sub-Nyquist rate. Prior to sampling, the analog signals are diversified using simple, implementable components. The diversification is achieved by injecting types of "structured randomness" into the ensemble, the result of which is subsampled. For reconstruction, the ensemble is modeled as a low-rank matrix that we have observed through an (undetermined) set of linear equations. Our main results show that this matrix can be recovered using standard convex programming techniques when the total number of samples is on the order of the intrinsic degree of freedom of the ensemble — the more heavily correlated the ensemble, the fewer samples are needed.

To motivate this study, we discuss how such ensembles arise in the context of array processing.

1 Introduction

This paper demonstrates how an ensemble of correlated signals can be efficiently sampled using three different implementable architectures. For each of these architectures we derive a sampling theorem that relates the bandwidth and the (a priori unknown) correlation structure to the total number of samples per second needed to fully capture the ensemble.

We consider ensembles of signals output from M sensors, each of which is bandlimited to frequencies below $W/2$ (see Figure 1). The entire ensemble can be acquired by taking W uniformly spaced samples per second in each channel, for a total sampling rate of MW. We will show that if the signals are correlated, meaning that the ensemble can be written as (or closely approximated by) distinct linear combinations of $R \ll M$ latent signals, then this net sampling rate can be reduced to the order of $\approx RW$ using *coded acquisition*. The sampling architectures we propose are blind to the correlation structure of the signals; this structure is discovered as the signals are reconstructed.

Each architecture involves a different type of *analog diversification* which ensures that the signals are sufficiently "spread out" so each point sample captures information about the ensemble. Ultimately, what is measured are not actual samples of the individual signals, but rather are different linear combinations that combine multiple signals and capture information over an interval of time. In Section 2.4.1, we will show that these samples can be expressed as linear measurements of a low-rank matrix. Over the course of one second, we want to acquire an $M \times W$ matrix comprised of samples of the ensemble taken at the Nyquist rate. The proposed sampling architecture produces a series of linear combinations of entries of this matrix. Conditions

[*] A. A. and J. R. are with the School of Electrical and Computer Engineering at Georgia Tech in Atlanta, Georgia. Email: alikhan@mit.edu, jrom@ece.gatech.edu. This work was supported by NSF grant CNS-0910592, ONR grant N00014-11-1-0459, and a grant from the Packard Foundation.

under which a low-rank matrix can be effectively recovered from an underdetermined set of linear measurements have been the object of intense study in the recent literature [1–4]; the mathematical contributions in this paper show how these conditions are met by systems with clear implementation potential.

Our motivation for studying these architectures comes from classical problems in array signal processing. In these applications, one or more "narrowband" signals are measured at multiple sensors at different spatial locations. While narrowband signals can have significant bandwidth, they are modulated up to a high carrier frequency, making them very heavily spatially correlated as they arrive at the array. This correlation, which we review in more detail in Section 1.2, can be systematically exploited for spatial filtering (beamforming), interference removal, direction-of-arrival estimation, and multiple source separation. These activities all depend on estimates of the inter-sensor correlation matrix, and the rank of this matrix can typically be related to the number of sources that are present.

Compressive sampling has been used in array processing in the past: sparse regularization was used for direction of arrival estimation [5–7] long before any of the "sub-Nyquist" sampling theorems started to make the theoretical guarantees concrete [8–10]. These results (along with more recent works including [11–13]), show how exploiting the structure of the array response in free space (for narrowband signals, this consists of samples of a superposition of a small number of sinusoids) can be used to either super-resolve the DOA estimate or reduce the number of array elements required to locate a certain number of sources. A single sample is associated with each sensor, and the acquisition complexity scales with the number of array elements.

In this paper, we exploit this structure in a different way. Our goal is to completely reconstruct the time-varying signals at all the array elements. The structure we impose on this ensemble is more general than the spatial spectral sparsity in this previous work; we ask that the signals are correlated in some a priori unknown manner. Our ensemble sampling theorems remain applicable even when the array response depends on the position of the source in a complicated way. Moreover, our reconstruction algorithms are indifferent to what the spatial array response actually is, as long as the narrowband signals remain sufficiently correlated.

The paper is organized as follows. In Sections 1.1 and 1.2 we describe the signal model and its motivation from problems in array processing. In Section 1.3, we introduce the components (and their corresponding mathematical models) that we will use in our sampling architectures. In Section 2, we present the sampling architectures, show how the measurements taken correspond to generalized measurements of a low-rank matrix, and state the relevant sampling theorems. Numerical simulations, illustrating our theoretical results, are presented in Section 3. Finally, Section 5, and Section 8 provide the derivation of the theoretical results.

1.1 Signal model

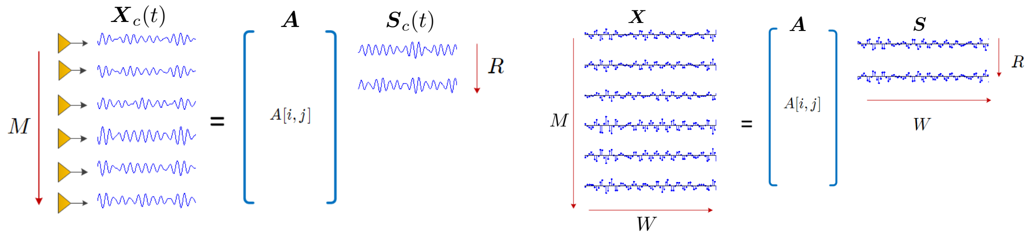
Our signal model is illustrated in Figure 1. We use $\boldsymbol{X_c}(t)$ to denote the continuous-time signal ensemble we are trying to acquire, and $x_1(t),\ldots,x_M(t)$ to denote the individual signals within that ensemble. Conceptually, we may think of $\boldsymbol{X}_c(t)$ as a "matrix" with a finite number M of rows, with each row containing a bandlimited signal. Our underlying assumption is that at every time t, the vector $\boldsymbol{X}_c(t)\in\mathbb{R}^M$ lies in a fixed subspace $\mathcal{S}$ of dimension R. We write

$$\boldsymbol{X}_c(t)\approx\boldsymbol{A}\boldsymbol{S}_c(t), \tag{1}$$

where $\boldsymbol{S}_c(t)$ is a smaller signal ensemble with R signals and $\boldsymbol{A}$ is a $M\times R$ matrix with entries $A[m,r]$ whose columns span $\mathcal{S}$. We will use the convention that fixed matrices operating to the left of the signal ensembles simply "mix" the signals point-by-point, and so (1) is equivalent to

$$x_m(t)\approx\sum_{r=1}^{R}A[m,r]s_r(t).$$

The only structure we will impose on the individual signals is that they are real-valued, bandlimited, and periodic. With this model, the signals live in a finite-dimensional linear subspace and there is a natural way



(a) (b)

Figure 1: *(a) Our model is that an ensemble of continuous-time signals are* correlated, *meaning the M signals can be closely approximated by a linear combination of R underlying signals. We can write the M signals in $\boldsymbol{X}_c(t)$ as a tall matrix (capturing the correlation structure) multiplied by an ensemble of R latent signals. (b) The matrix of samples inherits the low-rank structure of the continuous-time ensemble.*

to discretize the problem; that is, what exists in $\boldsymbol{X}_c(t)$ for $t \in [0,1]$ is all there is to know, and each signal can be captured exactly with W equally-spaced samples, which, for the most part, reduces the clutter in mathematics. Each bandlimited, periodic signal in the ensemble can be written as

$$x_m(t) = \sum_{k=-B}^{B} \alpha_m[k]\, e^{j2\pi kt}, \tag{2}$$

where the $\alpha_m[k]$ are complex but are symmetric, $\alpha_m[-k] = \alpha_m[k]^*$, to ensure that $x_m(t)$ is real. We can capture $x_m(t)$ perfectly by taking $W = 2B+1$ equally spaced samples per row. We will call this the $M \times W$ matrix of samples $\boldsymbol{X}$; knowing every entry in this matrix is the same as knowing the entire signal ensemble. We can write

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{F}^{\mathrm{H}}, \tag{3}$$

where $\boldsymbol{F}$ is a $W \times W$ normalized discrete Fourier matrix with entries

$$F[k,n] = \frac{1}{\sqrt{W}} e^{-j2\pi kn/W}, \quad 0 \le n \le W-1, \quad -B \le k \le B,$$

and $\boldsymbol{C}$ is a $M \times W$ matrix whose rows contain Fourier series coefficients for the signals in $\boldsymbol{X}_c(t)$. $\boldsymbol{F}$ is orthonormal, while $\boldsymbol{C}$ inherits the correlation structure of the ensemble $\boldsymbol{X}_c(t)$.

Same correlated signal model was considered in [14] for compressive sampling of multiplexed signals. Two multiplexing architectures were proposed and for each a sampling theorem was proved that dictated minimum number of samples for exact recovery of the signal ensemble. This paper presents sampling architectures, where we use a separate ADC for each channel and rigorously prove that ADCs can operate at roughly the optimal sampling rate to guarantee signal recovery. Other types of correlated signal models have been exploited previously to achieve gains in the sampling rate. For example, [15] shows that two signals related by a sparse convolution kernel can be reconstructed jointly at a reduced sampling rate. The signal model in [16] considers multiple signals all of which live in a fixed subspace spanned by a subset of the basis functions of a known basis, and provides some abstract results that show that the sampling rate required to successfully recover the signals scales with the number of basis functions used in the construction of the signals. In this paper, we also show that the sampling rate scales with the number of independent latent signals but we do this without the knowledge of the basis. For a more applied treatment of the results with similar flavor as in [16], we refer the reader to [17–19].

As will be shown later, we observe the signal ensemble $\boldsymbol{X}_c(t)$ through a limited set of random projections, and recover the latent signals, and the subspace from these few random projections by solving a nuclear norm minimization program. A related work [20] considers the case when given a few random projections of a signal, we find out the subspace to which it belongs by solving a series of least-squares programs.

We end this section by noting that their are many ways this problem might be discretized. Using Fourier series is convenient in two ways: we can easily tie together the notion of a signal being bandlimited with having a limited support in Fourier space, and our sampling operators have representations in Fourier space that make them more straightforward to analyze. In practice, however, we might choose to represent the signal over a finite interval using any one of a number of basis expansions — the low rank structure is preserved under any linear representation. It is also possible that we are interested in performing the ensemble recovery over multiple time frames, and would like the recovery to transition smoothly between these frames. For this we might consider a windowed Fourier series representations (e.g. the lapped orthogonal transform in [21]) that are carefully designed so that the basis functions are tapered sinusoids (so we again get something close to bandlimited signals by truncating the representation to a certain depth) but remain orthonormal. It is also possible to adjust our recovery techniques to allow for measurements which span consecutive frames, yielding another natural way to tie the reconstructions together.

1.2 Applications in array signal processing

One application area where low-rank ensembles of signals play a central role is array processing of narrowband signals. In this section, we briefly review how these low-rank ensembles arise. The central idea is that sampling a wavefront at multiple locations in space (as well as in time) leads to redundancies which can be exploited for spatial processing. These concepts are very general, and are common to applications as diverse as surveillance radars, underwater acoustic source localization and imaging, seismic exploration, wireless communications.

The essential scenario is that multiple signals are emitted from different locations, each of the signals occupies the same bandwidth of size W which has been modulated up to a carrier frequency f_c. The signals observed by receivers in the array are, to a rough approximation, complex multiples of one another. To a very close approximation, the observed signals lie in a subspace with dimension close to one — this subspace is determined by the location of the source. This redundancy between the observations at the array elements is precisely what causes the ensemble of signals to be low rank; the rank of the ensemble is determined by the number of emitters. The only conceptual departure from the discussion in previous sections, as we will see below, is that each emitter may be responsible for a subspace spanned by a number of latent "signals" that is greater than one (but still small).

Having an array with a large number of appropriately spaced elements can be very advantageous even when there only a relatively small number of emitters present. Observing multiple delayed versions of a signal allows us to perform spatial processing, we can beamform to enhance or null out emitters at certain angles, and separate signals coming from different emitters. The resolution to which we can perform this spatial processing depends on the number of elements in the array (and their spacing).

The main results of this paper do not give any guarantees about how well these spatial processing tasks can be performed. Rather, they say that the same correlation structure that makes these tasks possible can be used to lower the net sampling rate over time. The entire signal ensemble can be reconstructed from this reduced set of samples, and spatial processing can follow.

We now discuss in more detail how these low rank ensembles come about. For simplicity, this discussion will center on linear arrays in free space. As we just need the signal ensemble to lie in a low dimensional subspace, and do not need to know what this subspace may be beforehand, the essential aspects of the model extend to general array geometries channel responses and frequency-selective/multipath channels.

Suppose that a signal is incident on the array (as a plane wave) at an angle θ. Each array element observes a different shift of this signal — if we denote what is seen at the array center (the origin in Figure 2(a)) by $s(t)$, then an element m at distance d_m from the center sees $x_m(t) = s(t - (d_m/c)\sin\theta)$. If the signal consists of a single complex sinusoid, $s(t) = e^{j2\pi ft}$, then these delays translated into different (complex) linear multiples of the same signal,

$$x_m(t) = e^{-j2\pi f r_m \sin(\theta)/c}\, e^{j2\pi ft}. \tag{4}$$

In this case, the signal ensemble has rank[1] 1; we can write $\boldsymbol{X}(t) = \boldsymbol{a}(\theta, f)e^{j2\pi ft}$, where $\boldsymbol{a}(\theta, f)$ is an M-dimensional *steering* vector of complex weights given above.

This decomposition of the signal ensemble makes it clear how spatial information is coded into the array observations. For instance, standard techniques [22,23] for estimating the direction of arrival involve forming the spatial correlation matrix by averaging in time,

$$\boldsymbol{R}_{xx} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}(t_n)\boldsymbol{X}(t_n)^{\mathrm{H}}.$$

As the column space of $\boldsymbol{R}_{xx}$ should be $\boldsymbol{a}(\theta, f)$, we can correlate the steering vector for every direction to see which one comes closest to matching the principal eigenvector of $\boldsymbol{R}_{xx}$.

The ensemble remains low rank when the emitter has a small amount of bandwidth relative to a larger carrier frequency. If we take $s(t) = s_b(t)\, e^{j2\pi f_c t}$, where $s_b(t)$ is bandlimited to $W/2$, then when $W \ll f_c$, the $\boldsymbol{a}(\theta, f)$ for $f \in [f_c - W/2, f_c + W/2]$ will be very closely correlated with one another. In the standard scenario where the array elements are uniformly spaced $c/(2f_c)$ along a line, we can make this statement more precise using classical results on spectral concentration [24, 25]. In this case, the steering vectors $\boldsymbol{a}(\theta, f)$ for $f \in [f_c \pm W/2]$ are equivalent to integer spaced samples of a signal whose (continuous-time) Fourier transform is bandlimited to frequencies in $(1 \pm W/(2f_c))(\sin\theta)/2$, for a bandwidth less than $W/(2f_c)$. Thus the dimension of the subspace spanned by $\{\boldsymbol{a}(\theta, f),\ f \in [f_c \pm W/2]\}$ is, to within a very good approximation, $\approx MW/f_c + 1$.

Figure 2(b) illustrates a particular example. The plot shows the (normalized) eigenvalues of the matrix

$$\boldsymbol{R}_{aa} = \int_{f_c - W/2}^{f_c + W/2} \boldsymbol{a}(\theta, f)\boldsymbol{a}(\theta, f)^{\mathrm{H}}\, df, \tag{5}$$

for the fixed values of $f_c = 5$ GHz, $W = 100$ MHz, c equals the speed of light, $M = 101$, and $\theta = \pi/4$. We have $MW/f_c + 1 = 3.02$, and only 3 of the eigenvalues are within a factor of 10^4 of the largest one.

It is fair, then, to say that the rank of the signal ensemble is a small constant times the number of narrow band emitters.

1.3 Architectural components

In addition to analog-to-digital converters, our proposed architectures will use three standard components: analog vector-matrix multipliers, modulators, and linear time-invariant filters. The signal ensemble is passed through these devices, and the result is sampled using an analog-to-digital converter (ADC) taking either uniformly or non-uniformly spaced samples — these samples are the final outputs of our acquisition architectures.

The analog vector-matrix multiplier (AVMM) produces an output signal ensemble $\boldsymbol{A}\boldsymbol{X}_c(t)$ when we input it with signal ensemble $\boldsymbol{X}_c(t)$, where $\boldsymbol{A}$ is an $N \times M$ matrix whose elements are fixed. Since the matrix operates pointwise on the ensemble of signals, sampling output $\boldsymbol{A}\boldsymbol{X}_c(t)$ is the same as applying $\boldsymbol{A}$ to matrix $\boldsymbol{X}$ of the samples (i.e., sampling commutes with the application of $\boldsymbol{A}$). Recently, AVMM blocks have been built with hundreds of inputs and outputs and with bandwidths in the tens-to-hundreds of megahertz [26,27]. We will use the AVMM block to ensure that energy disperses more or less evenly throughout the channels. If $\boldsymbol{A}$ is a random orthogonal transform, it is highly probable that each signal in $\boldsymbol{A}\boldsymbol{X}_{\boldsymbol{c}}(t)$ will contain about the same amount of energy regardless of how the energy is distributed among the signals in $\boldsymbol{X}_{\boldsymbol{c}}(t)$ (formalized in Lemma 1 below), allowing us to deploy equal sampling resources in each channel while ensuring that resources on channels that are "quiet" are not being wasted.

[1] We are using complex numbers here to make the discussion go smoothly; the real part of the signal ensemble is rank 2, having a $\cos(2\pi ft)$ and a $\sin(2\pi ft)$ term.

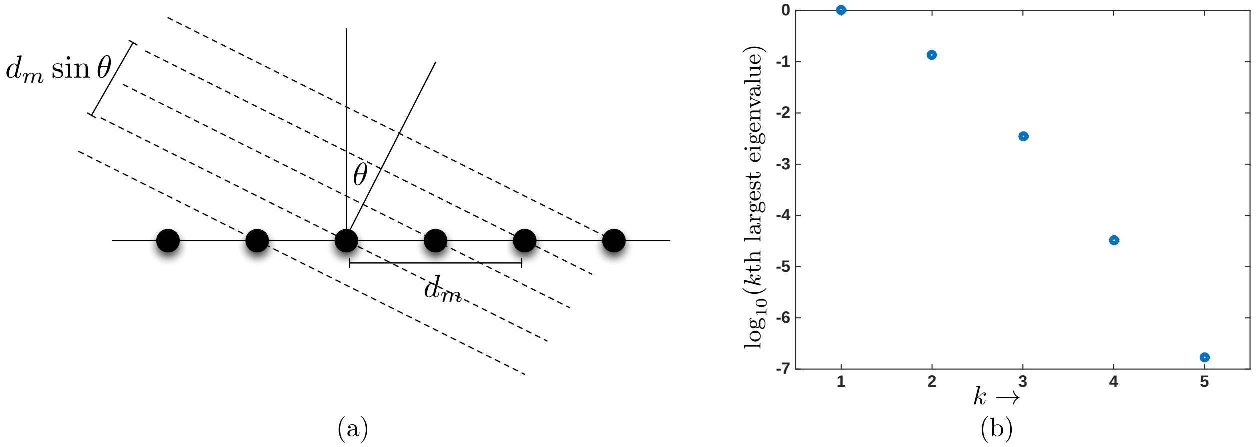


Figure 2: *(a) A plane wave impinges on a linear array in free space. When the wave is a pure tone in time, then the responses at each element will simply be phase shifts of one another. (b) Eigenvalues for $\boldsymbol{R}_{aa}$, on a $\log_{10}$ scale and normalized so that the largest eigenvalue is* 1*, defined in* (5) *for an electromagnetic signal width a bandwidth of* 100 *MHz and a carrier frequency of* 5 *GHz; the array elements are spaced half a carrier-wavelength apart. Even when the signal has an appreciable bandwidth, the signals at each of the array elements are heavily correlated — the effective dimension in this case is* $R = 3$ *or* 4*.*

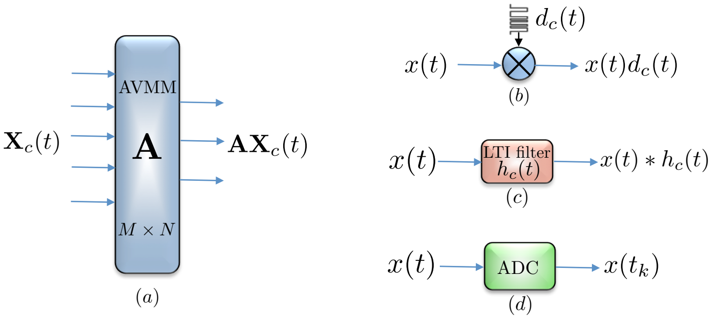


Figure 3: *(a) The analog vector-matrix multiplier (AVMM) takes random linear combinations of M input signals to produce N output signals. The action of AVMM can be thought of as the left multiplication of random matrix $\boldsymbol{A}$ to ensemble $\boldsymbol{X}_c(t)$. Intuitively, this operation amounts to distributing energy in the ensemble equally across channels. (b) Modulators multiply a signal in analog with a random binary waveform that disperses energy in the Fourier transform of the signal. (c) Random LTI filters randomize the phase information in the Fourier transform of a given signal by convolving it with $h_c(t)$ in analog, which distributes energy in time. (d) Finally, ADCs convert an analog stream of information in discrete form. We use both uniform and non-uniform sampling devices in our architectures.*

The second component of the proposed architecture is the modulators, which simply take a single signal $x(t)$ and multiply it by fixed and known signal $d_c(t)$. We will take $d_c(t)$ to be a binary ±1 waveform that is constant over time intervals of a certain length $1/W$. That is, the waveform alternates at the Nyquist sampling rate. If we take W samples of $d_c(t)x(t)$ on $[0,1]$, then we can write the vector of samples $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}, \tag{6}$$

where $\boldsymbol{x}$ is the W-vector containing the Nyquist-rate samples of $x(t)$ on $[0,1]$, and $\boldsymbol{D}$ is an $W \times W$ diagonal matrix whose entries are samples $\boldsymbol{d} \in \mathbb{R}^W$ of $d_c(t)$. We will choose a binary sequence that randomly generates $d_c(t)$, which amounts to $\boldsymbol{D}$ being a random matrix of the following form:

$$\boldsymbol{D} = \begin{bmatrix} d[0] & & & \\ & d[1] & & \\ & & \ddots & \\ & & & d[W-1] \end{bmatrix} \quad \text{where } d[n] = \pm 1 \text{ with probability } 1/2, \tag{7}$$

and the $d[n]$ are independent. Conceptually, the modulator disperses the information in the entire band of $x(t)$ — this allows us to acquire the information at a smaller rate by filtering a sub-band as will be shown in Section 2.

Compressive sampling architectures based on the random modulator have been analyzed previously in the literature [18, 28]. The principal finding is that if the input signal is spectrally sparse (meaning the total size of the support of its Fourier transform is a small percentage of the entire band), then the modulator can be followed by a low-pass filter and an ADC that takes samples at a rate comparable to the size of the active band. This architecture has been implemented in hardware in multiple applications [17, 29–32].

The third type of component we will use to preprocess the signal ensemble is a linear time-invariant (LTI) filter that takes an input $x(t)$ and convolves it with a fixed and known impulse response $h_c(t)$. We will assume that we have complete control over $h_c(t)$, even though this brushes aside admittedly important implementation questions. Because $x(t)$ is periodic and bandlimited, we can write the action of the LTI filter as a $W \times W$ circular matrix $\boldsymbol{H}$ operating on samples $\boldsymbol{x}$ (the first row of $\boldsymbol{H}$ consists of samples $\boldsymbol{h}$ of $h_c(t)$), that is, $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$, where $\boldsymbol{y}$ is the W Nyquist samples of the signal obtained at the output of the filter. We will make repeated use of the fact that $\boldsymbol{H}$ is diagonalized by the discrete Fourier transform:

$$\boldsymbol{H} = \boldsymbol{F}^{\mathrm{H}}\hat{\boldsymbol{H}}\boldsymbol{F}, \tag{8}$$

where $\boldsymbol{F}$ is the $W \times W$ normalized discrete Fourier matrix with entries, and $\hat{\boldsymbol{H}}$ is a diagonal matrix whose entries are $\hat{\boldsymbol{h}} = \sqrt{W}\boldsymbol{F}\boldsymbol{h}$. The vector $\hat{\boldsymbol{h}}$ is a scaled version of the non-zero Fourier series coefficients of $h_c(t)$.

To generate the impulse response, we will use a random unit-magnitude sequence in the Fourier domain . In particular, we will take

$$\hat{\boldsymbol{H}} = \begin{bmatrix} \hat{h}(0) & & & \\ & \hat{h}(1) & & \\ & & \ddots & \\ & & & \hat{h}(W-1) \end{bmatrix} \tag{9}$$

where

$$\hat{h}(\omega) = \begin{cases} \pm 1, \text{with prob. } 1/2, & \omega = 0 \\ e^{j\theta_\omega}, \text{ with } \theta_\omega \sim \text{Uniform}([0, 2\pi]), & 1 \le \omega \le (W-1)/2 \\ \hat{h}(W-\omega)^*, & (W+1)/2 \le \omega \le W-1 \end{cases} . \tag{10}$$

These symmetry constraints are imposed so that $\boldsymbol{h}$ (and hence, $h_c(t)$) is real-valued. Conceptually, convolution with $h_c(t)$ disperses a signal over time while maintaining fixed energy (note that $\boldsymbol{H}$ is an orthonormal matrix).

Convolution with a random pulse followed by sub-sampling has also been analyzed in the compressed sensing literature [33–36]. If the random filter is created in the Fourier domain as above, then following the filter with an ADC that samples at random locations produces a universally efficient compressive sampling architecture — the number of samples that we need to recover a signal with only S active terms at unknown locations in any fixed basis scales linearly in S and logarithmically in ambient-dimension W.

2 Main Results: Sampling Architectures

This section presents the main results and their implications for the problem of efficient sampling of the ensemble of correlated signals. We start with a straightforward architecture in Section 2.1 that minimizes the sample rate when the correlation structure is *known*. We then combine our components from the last section in different ways to create architectures that are provably effective under different assumptions on the signal ensemble. The first architecture in Section 2.3 consists simply of a bank of non-uniform ADCs; this architecture is provably effective when the ensemble is spread approximately evenly across channels and in time. The second architecture in Section 2.4 uses a random pre-modulator in front of a uniform ADC; this architecture is again effective when the energy in the ensemble is approximately uniform in both time and across the array. In Section 2.5, we present two architectures, consists of a AVMM, LTI filters, and either non-uniform ADCs or modulators coupled with uniform ADCs, that are *universal* in that they work for any type of correlation structure.

2.1 Fixed projections for known correlation structure

If the mixing matrix $\boldsymbol{A}$ for ensemble $\boldsymbol{X}_c(t)$ is known, then a straightforward way exists to sample the ensemble efficiently. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ be the singular value decomposition of $\boldsymbol{A}$, where $\boldsymbol{U}$ is $M \times R$ matrix with orthogonal columns, $\boldsymbol{\Sigma}$ is $R \times R$ diagonal matrix, and $\boldsymbol{V}$ is $W \times R$ with orthogonal columns. An efficient way is to *whiten* ensemble $\boldsymbol{A}$ with $\boldsymbol{U}^{\mathrm{T}}$ and sample the resulting R signals (each at rate W). This scheme is shown in Figure 4. $\boldsymbol{X}$ can be written as a multiplication of matrix $\boldsymbol{U}$ and $R \times W$ matrix $\boldsymbol{Y}$, which contains the Nyquist samples of signals $\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_R(t)$ respectively in its R rows. The signal ensemble can the be obtained using the multiplication

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{Y}.$$

Therefore, if we know the correlation structure $\boldsymbol{U}$, then $\boldsymbol{X}$ and hence $\boldsymbol{X}_c(t)$ (using sinc interpolation of samples in $\boldsymbol{X}$) can be recovered from the optimal total sampling rate of RW samples per second.

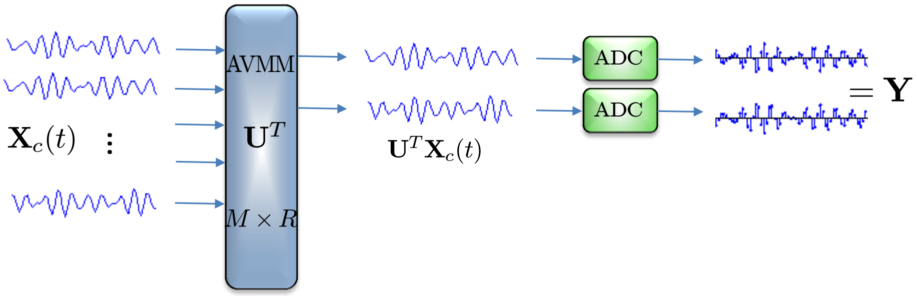


Figure 4: *If we know the correlation structure then efficient sampling structure is to* whiten *with $\boldsymbol{U}^{\mathrm{T}}$ and then sample, which requires R ADCs, each operating at a rate W samples per second. Hence, ADCs take a total of RW samples per second, RW being the degrees of freedom in the R signals bandlimited to $W/2$.*

In many interesting applications, the correlation structure of the ensemble $\boldsymbol{X}_c(t)$ is not known at the time of acquisition. In this paper, we design sampling strategies that are blind to the correlation structure but still achieve a near optimal sampling rate by introducing AVMMs, filters, and modulators. The randomness introduced by these components disperses energy over time and across channels so that the ADCs are always sensing information; this injection of structured randomness allows an efficient use of sampling resources.

2.2 Matrix recovery

Given the generalized samples of the ensemble, we recover (the discretized) $\boldsymbol{X}_0$ using a standard convex program. We denote the linear operator that maps the ensemble to the samples as $\mathcal{A}(\cdots)$, and our measurements as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_0), \quad \boldsymbol{y} \in \mathbb{R}^L, \quad \boldsymbol{X}_0 \in \mathbb{R}^{M \times W}.$$

To recover $\boldsymbol{X}_0$ from $\boldsymbol{y}$, we solve we solve

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_* \\ \text{subject to} \quad & \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}) \end{aligned} \tag{11}$$

where $\|\boldsymbol{X}\|_*$ is the nuclear norm (the sum of the singular values of $\boldsymbol{X}$). Minimizing the nuclear norm encourages the solution to be low rank [1], and has concrete performance guarantees when $\mathcal{A}$ obeys certain properties [2]. When the measurements are noisy

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_0) + \boldsymbol{\xi}, \quad \text{with } \|\boldsymbol{\xi}\|_2 \leq \delta \tag{12}$$

we instead solve the following quadratically constrained convex optimization program

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_* \\ \text{subject to} \quad & \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2 \leq \delta. \end{aligned} \tag{13}$$

Again, this relaxed program has certain performance guarantees [37, 38] when $\boldsymbol{X}_0$ is indeed low rank.

Our results will relate the total number of samples L taken by each architecture to the rank (and other properties) of the signal ensemble. The goal is to come as close to RW total samples as possible, the number necessary even when the correlation structure is known.

2.3 Architecture 1: Random sampling of time-dispersed correlated signals

The architecture presented in this section, shown in Figure 5, consists of one non-uniform sampling (NUS) ADC per channel. Each ADC takes samples at randomly selected locations, and these locations are chosen independently from channel to channel. Over the time interval $t \in [0, 1)$, a NUS ADC takes input signal $x_m(t)$ and returns the samples $\{x_m(t_k) : t_k \in T_m \subset \{0, 1/W, \ldots, 1 - 1/W\}$. The average sampling rate in each channel is $|T_m| = \Omega$. The net action of all M NUS ADCs is to return $M\Omega$ random samples of the input signal ensemble on uniform grid.

The sampling model is equivalent to observing $L = M\Omega$ randomly chosen entries of the matrix of samples $\boldsymbol{X}$, defined in (3). This problem is exactly the same as the matrix-completion problem [3], which states that given a small number of entries of a low-rank matrix, we can *fill in* the missing entries under some incoherence assumptions on the matrix $\boldsymbol{X}$. Since $\boldsymbol{X}$ is rank-R, its svd is

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{14}$$

where $\boldsymbol{U} \in \mathbb{R}^{M \times R}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{R \times R}$, and $\boldsymbol{V} \in \mathbb{R}^{W \times R}$. The coherence is then defined as

$$\mu_0^2 = \max\left\{\frac{M}{R} \max_{1 \leq i \leq M} \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}_i\|_2^2, \frac{W}{R} \max_{1 \leq i \leq W} \|\boldsymbol{V}^{\mathrm{T}}\boldsymbol{e}_i\|_2^2, \frac{MW}{R}\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|_\infty^2\right\}. \tag{15}$$

Now the matrix-completion results [39] in the noiseless case assert that if

$$M\Omega \gtrsim C\mu_0^2 R(W + M)\log^2(W),$$

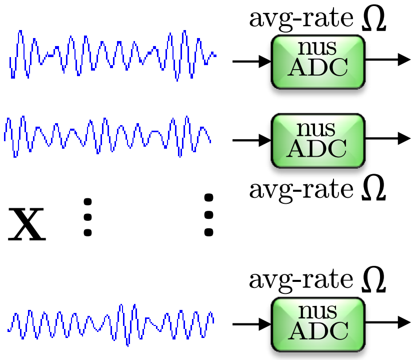


Figure 5: *M signals recorded by the sensors are sampled separately by the independent random sampling ADCs, each of which samples on a uniform grid at an average rate of Ω samples per second. This sampling scheme takes on the average a total of $M\Omega$ samples per second and is equivalent to observing $M\Omega$ entries of the matrix $\boldsymbol{X}$ at random*

then the solution of the nuclear norm minimization program (11) (with $\mathcal{A} : \mathbb{R}^{M\times W} \to \mathbb{R}^{M\Omega}$ such that $\mathcal{A}$ maps $\boldsymbol{X}$ to randomly chosen entries of $\boldsymbol{X}$) exactly equals $\boldsymbol{X}$ with high probability. The result indicates that the sampling rate scales (within some log factors) with the number R of independent signals rather than with the total number M of signals in the ensemble. When the measurements $\boldsymbol{y}$ are contaminated with noise as in (12) then the the result in [37] suggest that the solution $\tilde{\boldsymbol{X}}$ to the optimization problem (13) satisfies

$$\|\tilde{\boldsymbol{X}} - \boldsymbol{X}\|_{\mathrm{F}} \leq C_{\mu_0}\sqrt{\min(M,W)}\delta,$$

where C_{μ_0} is a constant that depends on the coherence μ_0, defined in (15).

As discussed before, the number of samples for the matrix completion increase with μ_0^2. The coherence parameter is small for matrices with even distribution of energy among their entries; see, [3] for details. Furthermore, signals are also known to be bandlimited, which implies that Architecture 1 is more effective for the efficient sampling of signals dispersed across channels and time. We will show in Section 2.5 that using AVMM and filters at the front end of the sampling scheme force the signal energy to be distributed evenly. This will allow us to build sampling architectures that are effective uniformly for all ensembles of signals lying in a subspace.

2.4 Architecture 2: The random modulator for correlated signals

To efficiently acquire the signal ensemble lying in a subspace, the architecture 2, shown in Figure 6, follows a two-step approach. In the first step, each of the M signals undergo analog preprocessing, which involves modulation, and low-pass filtering. The modulator takes an input signal $x_m(t)$ and multiplies it by a fixed and known $d_m(t)$. We will take $d_m(t)$ to be a binary ±1 waveform that is constant over an interval of length $1/W$. Intuitively, the modulation results in the diversification of the signal *information* over the frequency band of width W. The diversified analog signals are then processed by an analog-low-pass filter; implemented using an integrator, see [28] for details. The low-pass filter only selects a frequency sub-band (or a subspace) of width roughly equal to Ω, and as will be shown in Theorem 1, this partial information is enough for the signal reconstruction. The partial information suffices as the signals are scrambled using modulators before low-pass filtering. Note that the low-pass filter in each channel in Fig. 6 can be replaced; in general, by a band-pass filter, i.e., the location of the band does not matter only its width does. This also explains why we don't need to know the subspace in which signals live in advance.

In the second step, the filtered signal is sampled by an ADC in each channel at a lower rate Ω. The result in Theorem 1 asserts that Ω is roughly of a factor of R/M smaller than the Nyquist rate W.

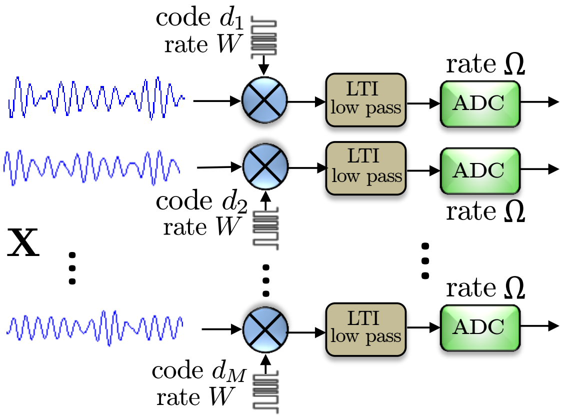


Figure 6: *The random demodulator for multiple signals lying in a subspace: M signals lying in a subspace are preprocessed in analog using a bank of independent modulators, and low-pass filters. The resultant signal is then sampled uniformly by an ADC in each channel operating at rate Ω samples per second. The net sampling rate is $L = \Omega M$ samples per second.*

2.4.1 System in matrix form

Each of the M input signals $x_m(t), 1 \leq m \leq M$ is multiplied by an independently generated random binary waveform $d_m(t), 1 \leq m \leq M$ alternating at rate W. That is, the output after the modulation in the mth channel is

$$y_m(t) = x_m(t) \cdot d_m(t), \quad 1 \leq m \leq M, \text{ and } t \in [0,1).$$

The $y_m(t)$ are then low-pass filtered using an integrator, which integrates $y_m(t)$ over an interval of width $1/\Omega$ and the result is then sampled at rate Ω using an ADC. The samples taken by the ADC in the mth channel are

$$y_m[n] = \int_{(n-1)/\Omega}^{n/\Omega} y_m(t)dt, \quad 1 \leq n \leq \Omega.$$

The integration operation commutes with the modulation process; hence, we can equivalently integrate the signals $x_m(t), 1 \leq m \leq M$ over the interval of width $1/W$, and treat them as samples $\boldsymbol{X}_0 \in \mathbb{R}^{M \times W}$ of the ensemble $\boldsymbol{X}_c(t)$. The entries $X_0[m,n]$ of the matrix $\boldsymbol{X}_0$ are

$$\begin{aligned} X_0[m,n] &= \int_{(n-1)/W}^{n/W} x_m(t)dt, \\ &= \sum_{\omega=-B}^{B} C[m,\omega] \left[\frac{e^{\iota 2\pi\omega/W} - 1}{\iota 2\pi\omega} \right] e^{-\iota 2\pi\omega n/W}, \end{aligned}$$

where the bracketed term representing the low-pass filter

$$\tilde{L}[\omega] = \left[\frac{e^{\iota 2\pi\omega/W} - 1}{\iota 2\pi\omega} \right]$$

is evaluated in the window $\omega \in \{-B, \dots, B\}$, where $W = 2B+1$, as defined in (2). We will use an equivalent evaluation $L[\omega]$ of $\tilde{L}[\omega]$ in the window $\omega \in \{1, \dots, 2B+1\}$, i.e., we will use $L[\omega] = \tilde{L}[\omega]$, for $\omega = 1, \dots, B$, and $L[\omega] = \tilde{L}[\omega - W + 1]$, for $\omega = B+1, \dots, 2B+1$. The Fourier coefficients of $C[m,\omega]$ of $\boldsymbol{X}$ defined in (3) are related to the Fourier coefficients $C_0[m,\omega]$ of $\boldsymbol{X}_0$

$$C_0[m,\omega] = C[m,\omega]L[\omega], \quad 1 \leq \omega \leq W, \tag{16}$$

and in time domain

$$\boldsymbol{X}_0 = \boldsymbol{C}_0 \boldsymbol{L} \boldsymbol{F}^{\mathrm{H}}, \tag{17}$$

where $\boldsymbol{L}$ is a $W \times W$ diagonal matrix containing $L[\omega]$, $1 \leq \omega \leq W$ as its diagonal entries, $\boldsymbol{F}$ is the $W \times W$ DFT matrix, and $\boldsymbol{C}_0$ is the coefficients matrix with entries defined in (16). Since $\boldsymbol{C}_0$ inherits its low-rank structure from $\boldsymbol{C}$; therefore, $\boldsymbol{X}_0$ is also a low-rank matrix of rank R. In the rest of this write up, we will consider recovering the rank R matrix $\boldsymbol{X}_0$. Since $\boldsymbol{L}$ is well-conditioned, the recovery of $\boldsymbol{X}_0$ implies the recovery of $\boldsymbol{X}$ in (3).

In light of (6), the W equally-spaced samples of $d_m(t)x_m(t)$ are $\boldsymbol{D}_m\boldsymbol{x}_m$, where $\boldsymbol{x}_m$ contains the W uniformly-spaced samples of $x_m(t)$, and $\boldsymbol{D}_m$, as in (7), is a random diagonal matrix containing random binary signs $d_m[n]$ along the diagonal. The binary waveform for the modulator in each channel is independently generated, which amounts to $\{\boldsymbol{D}_m\}$ being independent.

The samples $\boldsymbol{y}_m \in \mathbb{R}^{\Omega}$ in $t \in [0, 1)$ taken by the ADC in the mth branch are

$$\boldsymbol{y}_m = \boldsymbol{P}\boldsymbol{D}_m\boldsymbol{x}_m, \quad 1 \leq m \leq M,$$

where $\boldsymbol{x}_m \in \mathbb{R}^W$ are the rows of $\boldsymbol{X}_0$ defined in (17); $\boldsymbol{D}_m$ is the independent instantiation of $W \times W$ random diagonal matrix defined in (7), and corresponds to the modulator in the mth branch; and $\boldsymbol{P} : \Omega \times W$ is the matrix for the integrator (used as low-pass filter; for more details, see [28]) that contains ones in locations $(\alpha, \beta) \in (j, \mathcal{B}_j)$, for $1 \leq j \leq \Omega$, where

$$\mathcal{B}_j = \{(j-1)W/\Omega + 1 : jW/\Omega\}, \quad 1 \leq j \leq \Omega,$$

where we are assuming for simplicity that Ω is a factor of W. Since the action of the integrator commutes with the action of the modulator, the operation of the integrator can be simply represented as a block-diagonal matrix $\boldsymbol{P}$ operating on the modulated entries of the rows of $\boldsymbol{X}_0$, which contains the samples of the integrated signals. Putting it all together, the samples acquired by the ADCs can be written as a random block-diagonal matrix times the vector $\operatorname{vec}(\boldsymbol{X}_0)$, formed by stacking the rows of low-rank $\boldsymbol{X}_0$ as

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{P}\boldsymbol{D}_1 & & \\ & \ddots & \\ & & \boldsymbol{P}\boldsymbol{D}_M \end{bmatrix} \cdot \operatorname{vec}(\boldsymbol{X}_0), \tag{18}$$

where $\boldsymbol{y} \in \mathbb{R}^{\Omega M}$ is the vector containing the samples acquired by all the ADCs. We will denote by L, the total number of samples per second $M\Omega$ taken by all the ADCs.

2.4.2 Sampling theorem: Exact and stable recovery

Clearly, the samples $\boldsymbol{y}$ at the ADCs are a linear transformation $\mathcal{A}$ of the rank-R matrix $\boldsymbol{X}_0$

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_0).$$

Let $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ be the reduced form svd of $\boldsymbol{X}_0$ with $\boldsymbol{U} : M \times R$, $\boldsymbol{V} : W \times R$ being the matrices of left and right singular vectors, respectively, and $\boldsymbol{\Sigma} : R \times R$ being a diagonal matrix containing singular values of $\boldsymbol{X}_0$. Let $\{\boldsymbol{e}_i\}_{1 \leq i \leq M}$, and $\{\tilde{\boldsymbol{e}}_k\}_{1 \leq k \leq W}$ be the standard basis vectors of dimensions M, and W, respectively. The coherences of $\boldsymbol{X}_0$ is defined as

$$\mu_1^2 = \frac{M}{R} \max_{1 \leq i \leq M} \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}_i\|_2^2, \tag{19}$$

$$\mu_2^2 = \frac{W}{R} \max_{1 \leq j \leq W} \|\boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{e}}_j\|_2^2, \tag{20}$$

and

$$\mu_3^2 = \frac{M\Omega}{R} \max_{\substack{1 \leq i \leq M \\ 1 \leq j \leq \Omega}} \sum_{k \sim \mathcal{B}_j} \langle \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}, \boldsymbol{e}_i\tilde{\boldsymbol{e}}_k^{\mathrm{T}} \rangle^2. \tag{21}$$

A simple calculation shows that $1 \leq \mu_1^2 \leq M/R$, and $1 \leq \mu_2^2 \leq W/R$; see, [3] for details. We will only show here that $1 \leq \mu_3^2 \leq MW/R$. Begin with

$$\begin{aligned}\sum_{k\sim\mathcal{B}_j} \langle \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}, \boldsymbol{e}_i\tilde{\boldsymbol{e}}_k^{\mathrm{T}}\rangle^2 &\leq \sum_{k\sim\mathcal{B}_j} \max_i \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}_i\|_2^2 \cdot \max_k \|\boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{e}}_k\|_2^2 \\ &= |\mathcal{B}_j|\mu_1^2\mu_2^2\frac{R^2}{MW},\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz's inequality, and the last equality follows from the definitions in (19), and (20). The fact that $\mu_3^2 \leq MW/R$ follows by using the upper bounds on μ_1^2, and μ_2^2. Similarly, the lower bound is obtained by summing over j on both sides of the definition as follows

$$\sum_{j=1}^{\Omega}\mu_3^2 = \Omega\mu_3^2 \geq \frac{M\Omega}{R}\sum_{j=1}^{\Omega}\sum_{k\sim\mathcal{B}_j}\langle \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}, \boldsymbol{e}_i\tilde{\boldsymbol{e}}_k^{\mathrm{T}}\rangle^2 = \frac{M\Omega}{R}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}_i\|_2^2,$$

which immediately gives $\mu_3^2 \geq 1$ after using the fact using the fact that $\mu_1^2 \geq 1$. All three coherence quantities take smallest values for equally dispersed singular vectors and largest values for sparse singular vectors [3]. In our context, this implies that the coherence parameters quantify the dispersion of the signal-ensemble energy across time and channels.

Theorem 1. *Suppose $L = M\Omega$ measurements of the ensemble $\boldsymbol{X}_0$ are taken using* (18)*. If*

$$\Omega \geq C\beta\mu_3^2 R\max(\mu_1^2 W/M, \mu_2^2)\log^3(WM) \tag{22}$$

for some $\beta > 1$, then the minimizer $\tilde{\boldsymbol{X}}$ to the problem (11) *is unique and equal to $\boldsymbol{X}_0$ with probability at least $1-(WM)^{-\beta}$.*

The result indicates that each ADC operates at a rate Ω that is smaller than the Nyquist rate W by a factor of R/M. The net sampling rate L scales with the number R of independent signals rather than with the total number M of signals in the ensemble. Thus, the random demodulator provably acquires multiple signals lying in a subspace at a rate that is within log factors of the optimal sampling rate without knowing the subspace in advance. The coherence terms suggest that the sampling architecture is more effective for sampling signals with energy dispersed evenly across channels and time.

2.4.3 Stable recovery

In a realistic scenario, the measurements are almost always contaminated with noise, as in (12). In the case, when the noise is bounded, i.e., $\|\boldsymbol{\xi}\|_2 \leq \delta$, then following the template of the proof in [37], it can be shown that under the conditions of Theorem 1, the solution $\tilde{\boldsymbol{X}}$ of (13) obeys

$$\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathrm{F}} \leq C\sqrt{\min(W,M)}\delta, \tag{23}$$

with high probability; for more details on this, see a similar stability result in Theorem 2 in [40]. The above stability result is weak due to the multiplication factor $\sqrt{\min(W,M)}$. In contrast to the optimization program in (13), the solution $\tilde{\boldsymbol{X}}$ to a slightly different optimization program:

$$\tilde{\boldsymbol{X}} = \operatorname{argmin}_{\boldsymbol{X}}\{\|\boldsymbol{X}\|_{\mathrm{F}}^2 - 2\langle \boldsymbol{y}, \mathcal{A}(\boldsymbol{X})\rangle + \lambda\,\|\boldsymbol{X}\|_*\}, \tag{24}$$

proposed in [41] can be theoretically shown to obey essentially optimal stable recovery results. By completing the square, it is easy to see that the above estimator is equivalent to

$$\tilde{\boldsymbol{X}} = \operatorname{argmin}_{\boldsymbol{X}}\{\|\boldsymbol{X} - \mathcal{A}^*(\boldsymbol{y})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{X}\|_*\}.$$

Taking the sub-differential $\partial C(\boldsymbol{X})$ of the cost function $C(\boldsymbol{X}) = \{\|\boldsymbol{X} - \mathcal{A}^*(\boldsymbol{y})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{X}\|_*\}$ and using the fact $\tilde{\boldsymbol{X}}$ is the minimizer if and only if $\mathbf{0} \in \partial C(\boldsymbol{X})$, it can be shown [41] that the estimate $\tilde{\boldsymbol{X}}$ is a soft thresholding of the singular values of the matrix $\boldsymbol{X}_{\mathcal{A}} = \mathcal{A}^*(\boldsymbol{y}) \in \mathbb{R}^{M\times W}$

$$\tilde{\boldsymbol{X}} = \sum_i \{\sigma_i(\boldsymbol{X}_{\mathcal{A}}) - \frac{\lambda}{2}\}_+ \boldsymbol{u}_i(\boldsymbol{X}_{\mathcal{A}})\boldsymbol{v}_i^*(\boldsymbol{X}_{\mathcal{A}}),$$

where $x_+ = \max\{x, 0\}$; in addition, $\boldsymbol{u}_i(\boldsymbol{X}_{\mathcal{A}})$, and $\boldsymbol{v}_i(\boldsymbol{X}_{\mathcal{A}})$ are the left and right singular vectors of the matrix $\boldsymbol{X}_{\mathcal{A}}$, respectively; and $\sigma_i(\boldsymbol{X}_{\mathcal{A}})$ is the corresponding singular value.

In comparison to the estimator (24), the matrix Lasso in (13) does not use the knowledge of the known distribution of $\mathcal{A}$ and instead minimizes the empirical risk $\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2 = \|\boldsymbol{y}\|_2^2 - 2\langle \boldsymbol{y}, \mathcal{A}(\boldsymbol{X})\rangle + \|\mathcal{A}(\boldsymbol{X})\|_2^2$. Knowing the distribution, and the fact that $\mathrm{E}\,\mathcal{A}^*\mathcal{A} = \mathcal{I}$ holds in our case, we replace $\|\mathcal{A}(\boldsymbol{X})\|_2^2$, by its expected value $\mathrm{E}\,\|\mathcal{A}(\boldsymbol{X})\|_2^2 = \|\boldsymbol{X}\|_{\mathrm{F}}^2$ in the empirical risk to obtain the estimator in (24). Although the KLT estimator is easier to analyze, and gives optimal stable recovery results in theory but it does not empirically perform as well as matrix Lasso.

Before stating the stable recovery results, we introduce the statistical assumptions on the additive measurement noise $\boldsymbol{\xi}$, which are given as

$$\max_{ij} \mathrm{E}\exp\left(\frac{|\xi_{ij}|^2}{\sigma^2}\right) < \tilde{c} \tag{25}$$

$$\|\boldsymbol{\xi}\|_{\psi_2}^2 = c\sum_{i,j} \mathrm{E}\,\xi_{ij}^2 = cL\sigma^2, \tag{26}$$

where ψ_2 denotes the Orlicz-2 norm of vector $\boldsymbol{\xi} \in \mathbb{R}^L$ that contains ξ_{ij} as its entries. The choice of the indexing with double-index i, j will be clear in Section 5. With this the following result is in order.

Theorem 2. *Let $\boldsymbol{X}_0 \in \mathbb{R}^{M\times W}$ be a rank R matrix, and suppose that we observe y_{ij} as in* (12) *contaminated with noise ξ_{ij} such that* (25) *holds. Then with probability at least $1 - (WM)^{-\beta}$ for some $\beta > 1$, the solution $\tilde{\boldsymbol{X}}$ to* (24) *will obey*

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathrm{F}} \leq C\|\boldsymbol{\xi}\|_{\psi_2}, \tag{27}$$

for a fixed constant C, when $\Omega \geq C\beta\mu_3^2 \max(W/M, 1)\log^2(WM)$

Roughly speaking, the stable recovery theorem states that the nuclear norm penalized estimators are stable in the presence of additive measurement noise. The results in Theorem 2 are derived assuming that ξ_{ij} are random with statistics in (25). In contrast, the stable recovery results in the compressed sensing literature only assume that the noise is bounded, i.e., $\|\boldsymbol{\xi}\|_2 \leq \delta$, where $\boldsymbol{\xi}$ is the noise vector introduced earlier. Here, we give a brief comparison of the results in Theorem (2) with the stable recovery results in [37, 42].

Compare the result in (23) with (27), it follows that our results improve upon the results in [37] by a factor of $1/(W \wedge M)$. We will also compare our stable recovery results against the stable recovery results derived in [42]. The result roughly states if the linear operator $\mathcal{A}$ satisfies the matrix RIP [2], and $\|\boldsymbol{\xi}\|_2 \leq \delta$, then the solution $\tilde{\boldsymbol{X}}$ to (13) obeys

$$\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathrm{F}} \leq C\delta. \tag{28}$$

The above result is essentially optimal stable recovery result. In comparison to (28), the result in (27) is also optimal, however, we prove it for a different estimator and under a statistical bound on the noise term $\|\boldsymbol{\xi}\|_{\psi_2} \leq \delta$. In addition, we also donot require the matrix RIP for $\mathcal{A}$, which is generally required to prove optimal results of the form of (28).

The result in Theorem 1 is more effective for incoherent $\boldsymbol{X}$. Roughly speaking, the incoherence conditions are satisfied by a matrix with even distribution of energy among its entries. The incoherence conditions on the matrix of samples $\boldsymbol{X}_0$ when combined with the fact that the signals are also known to be bandlimited implies that Architectures 2 is also feasible for the efficient sampling of *spread out* correlated signal ensembles. In contrast, a *universal* sampling scheme will work for any ensemble of correlated signals. We can design such a sampling scheme by preprocessing signals in analog by components that transform (with high probability) matrix $\boldsymbol{X}$ to an incoherent matrix. We present such an architecture in the following section.

2.5 Architecture 3: Uniform sampling architectures

The performance of Architecture 1 and 2 depends on the coherences in defined (15); and (19), (20), (21) of ensemble $\boldsymbol{X}_c(t)$. That is, the net sampling rate depends on the energy distribution of the signal ensemble. In this section, we present uniform sampling architectures that sample any given ensemble of correlated signals with no prior requirements on the energy distribution. We will present *uniform* versions of Architecture 1 and 2 shown in Figure 7, and 8. In both uniform-sampling schemes, we force the coherences to be small by adding a little analog preprocessing using AVMM and filters at the front end of the Architecture 1 and 2.

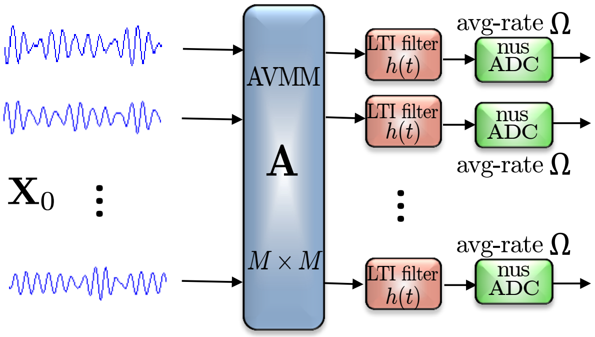


Figure 7: *Analog vector-matrix multiplier (AVMM) takes random linear combinations of M input signals to produce M output signals. This equalizes energy across channels. The random LTI filters convolve the signals with a diverse waveform that results in dispersion of signals across time. The resultant signals are then sampled, at locations selected randomly on a uniform grid, at an average rate Ω, using a non-uniform sampling (NUS) ADC in each channel.*

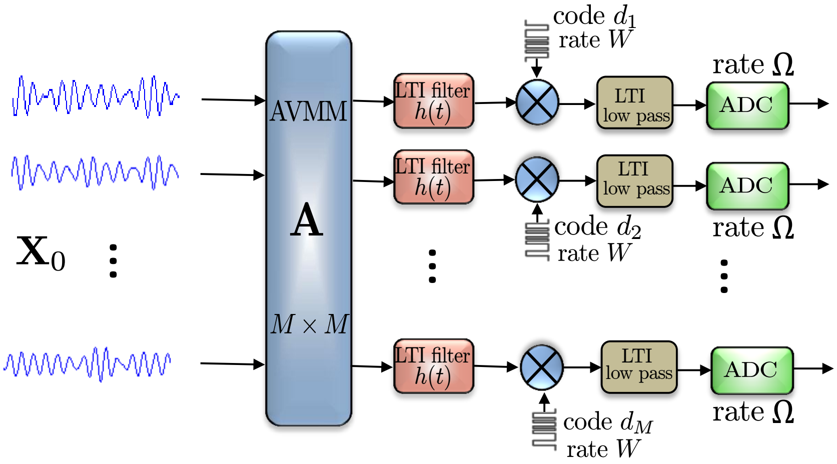


Figure 8: *Analog vector-matrix multiplier (AVMM) takes random linear combinations of M input signals to produce M output signals. This equalizes energy across channels. The random LTI filters convolve the signals with a diverse waveform that results in dispersion of signals across time. The resultant signals are then sampled uniformly at rate Ω using the random demodulator in each channel.*

The sampling architectures shown in Figure 7, and 8 preprocess the signals in analog with an analog-vector-matrix multiplier that spreads energy across channels. The analog ensemble is then processed by a bank of random filters that spread the energy over time. The combined action of the AVMM with a random matrix $\boldsymbol{A}$ and the analog LTI filters with a random matrix $\boldsymbol{H}$ forces the processed output $\boldsymbol{X}_\mathfrak{p}$ to be incoherent w.h.p. The incoherent signals are then either sampled randomly with an NUS ADC in each channel, as in

Architecture 1, or sampled uniformly using a modulator, an integrator, and a uniform ADC in each channel, as in Architecture2.

The AVMM takes the random linear combination of M input signals to produce M output signals. The action of the AVMM can be modeled by left multiplication of random matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ with ensemble $\boldsymbol{X}$, which then equalizes w.h.p., the energy in each of the channels regardless of the initial energy distribution. Furthermore, the all pass LTI filters convolve the signals with a diverse impulse response $h_c(t)$, which disperses signal energy over time w.h.p. (see Lemma 1). We will use the same random LTI filter $h_c(t)$ in each channel. The action of the random convolution [33] of $h_c(t)$ with each signal in the ensemble can be modeled by the right multiplication of a circulant random orthogonal matrix $\boldsymbol{H} \in \mathbb{R}^{W\times W}$ with $\boldsymbol{X}$, assuming W is even; it will be clear how to extend the argument to W odd. We can write $\boldsymbol{H} = \boldsymbol{W}\boldsymbol{Q}^*$, where

$$\boldsymbol{Q}[n,\omega] = \begin{cases} \frac{1}{\sqrt{W}} & \omega = 0 \\ \frac{2}{\sqrt{W}}\cos\left(\frac{2\pi\omega n}{W}\right) & \omega = [1, \frac{W}{2}-1] \\ \frac{1}{\sqrt{W}}(-1)^{k-1} & \omega = \frac{W}{2} \\ \frac{2}{\sqrt{W}}\sin\left(\frac{2\pi\omega n}{W}\right) & \omega = [\frac{W}{2}+1, W-1] \end{cases} \tag{29}$$

$$\boldsymbol{W}[n,\omega] = \begin{cases} \frac{z_0}{\sqrt{W}}, & \omega = 0 \\ \frac{2}{\sqrt{W}}\cos\left(\frac{2\pi\omega n}{W} + \theta_\omega\right) & \omega = [1, \frac{W}{2}-1] \\ \frac{z_{W/2}}{\sqrt{W}}(-1)^{k-1}, & \omega = \frac{W}{2} \\ \frac{2}{\sqrt{W}}\sin\left(\frac{2\pi\omega n}{W} + \theta_\omega\right), & \omega = [\frac{W}{2}+1, W-1] \end{cases} . \tag{30}$$

and $z_0, z_{W/2} = \pm 1$ with equal probability and θ_ω for $\omega = 1, \ldots, W/2-1$ are uniform on $[0, 2\pi]$ and all $W/2+1$ of these random variables are independent.

Application of the AVMM with random orthogonal $\boldsymbol{A}$ and the LTI random filters with random orthogonal $\boldsymbol{H}$ on input ensemble $\boldsymbol{X}$ spreads signals out across channels and over time w.h.p. As a result, we obtain $\boldsymbol{X}_{\mathtt{p}} \in \mathbb{R}^{M\times W}$:

$$\boldsymbol{X}_{\mathtt{p}} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}. \tag{31}$$

Let $\boldsymbol{U}_{\mathtt{p}} = \boldsymbol{A}\boldsymbol{U}$ and $\boldsymbol{V}_{\mathtt{p}} = \boldsymbol{H}\boldsymbol{V}$, where $\boldsymbol{U}_{\mathtt{p}} \in \mathbb{R}^{M\times R}, \boldsymbol{V}_{\mathtt{p}} \in \mathbb{R}^{W\times R}$ be the left and right singular vectors of matrix $\boldsymbol{X}_{\mathtt{p}}$, respectively. Note that matrix $\boldsymbol{X}_{\mathtt{p}}$ is an isometry with $\boldsymbol{X}$ and has the same rank as $\boldsymbol{X}$. The left and right singular vectors $\boldsymbol{U}_{\mathtt{p}}$ and $\boldsymbol{V}_{\mathtt{p}}$ of $\boldsymbol{X}_{\mathtt{p}}$ are in some sense random orthogonal matrices and hence, incoherent w.h.p. The following Lemma shows the incoherence of matrix $\boldsymbol{X}_{\mathtt{p}}$.

Lemma 1. *Fix matrices $\boldsymbol{U} \in \mathbb{R}^{M\times R}$ and $\boldsymbol{V} \in \mathbb{C}^{W\times R}$ of the left and right singular vectors, respectively, each of which consists of R orthogonal unit norm columns. Create random orthonormal matrices $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ and $\boldsymbol{H} \in \mathbb{R}^{W\times W}$. Then*

- $\max_{1\le i\le M} \|\boldsymbol{U}_{\mathtt{p}}^*\boldsymbol{e}_i\|_2^2 \le C_\beta \max(R, \log M)/M$ *with a probability exceeding* $1 - M^{-\beta}$.
- $\max_{1\le j\le W} \|\boldsymbol{V}_{\mathtt{p}}^*\boldsymbol{e}_j\|_2^2 \le C_\beta \max(R, \log W)/W$ *with a probability exceeding* $1 - W^{-\beta}$.
- $\max_{\substack{1\le i\le M\\1\le j\le W}} \langle \boldsymbol{U}_{\mathtt{p}}\boldsymbol{V}_{\mathtt{p}}^*, \boldsymbol{e}_i\tilde{\boldsymbol{e}}_j^*\rangle^2 \le C_\beta \log W \max(R, \log M)/MW$ *with a probability exceeding* $1 - O(W^{-\beta} + M^{-\beta})$.
- $\max_{\substack{1\le i\le M\\1\le j\le \Omega}} \sum_{k\sim\mathcal{B}_j} \langle \boldsymbol{U}_{\mathtt{p}}\boldsymbol{V}_{\mathtt{p}}^*, \boldsymbol{e}_i\tilde{\boldsymbol{e}}_k\rangle^2 \le C_\beta \log W \max(R, \log M)/M\Omega$ *with a probability exceeding* $1 - O(W^{-\beta} + M^{-\beta})$.

Proof of Lemma 1 is presented in Section 4.

2.5.1 Sufficient sampling rate for the first uniform sampling architecture

Lemma 1 combined with the definition (15) shows that the coherence parameter $\mu_0^2 \leq C_\beta \log(W)$ holds for for $R > \log M$ with high probability. Using this bound in the matrix-completion results [3,37] in the noiseless case asserts that if $M\Omega \gtrsim CR(W+M)\log^3(W)$, the solution of the nuclear norm minimization program (11) exactly equals $\boldsymbol{X}$ with high probability. We are paying an extra log factor in the measurements but now there is no dependence on the energy distribution of the ensemble.

2.5.2 Sufficient sampling rate for the second uniform sampling architecture

Combining Lemma 1 with Theorem 1 immediately provides with the following corollary.

Corollary 1. *Suppose Ω measurements of the ensemble $\boldsymbol{X}_0$ are taken through the uniform random demodulator setup. If*

$$\Omega \geq C\beta R \max(W/M, 1)\log^4(WM) \tag{32}$$

for some $\beta > 1$, and $R > \log M$, the minimizer $\tilde{\boldsymbol{X}}$ to the problem (11) *is unique and equal to $\boldsymbol{X}_0$ with probability at least $1 - O(WM)^{-\beta}$.*

Hence, we can recover $\tilde{\boldsymbol{X}}$ and hence, $\boldsymbol{X}$ in both uniform sampling architectures in Figure 7, and 8 using the nuclear-norm minimization.

3 Numerical Experiments

In this section, we study the performance of the proposed sampling architectures with some numerical experiments. Since the first sampling architecture reduces to an already well-studied matrix completion problem, we have chosen here to present only the sampling performance of Architecture 2, which reduces to a matrix recovery problem from a block-diagonal measuremnt matrix that has not been studied before.

3.1 Sampling performance

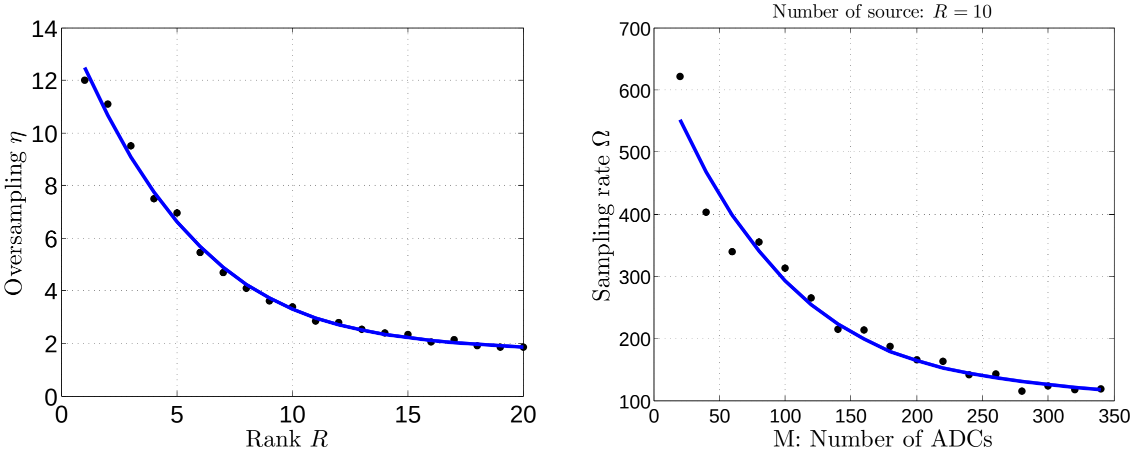
In all of the experiments in this section, we generate the unknown rank-R matrix $\boldsymbol{X}_0$ by the multiplication of a tall $M \times R$, and a fat $R \times W$ Gaussian matrices. Our objective is to recover a batch of $M = 100$ signals, with $W = 1024$ samples taken in a given window of time using Architecture 2. We will use the following parameters to evaluate the performance of the sampling architecture:

$$\text{Oversampling factor} : \eta = \frac{M\Omega}{R(W+M-R)},$$

where the oversampling factor is the ratio between the combined sampling rate of all the ADCs in Figure 6, and the degrees of freedom in rank-R matrix of samples $\boldsymbol{X}_0$. The successful reconstruction is declared when the relative error obeys

$$\text{Relative error} := \frac{\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\mathrm{F}}}{\|\boldsymbol{X}_0\|_{\mathrm{F}}} \leq 10^{-2}.$$

The first experiment shows a graph, in Figure 9(a), between the oversampling factor η, and R. Each point, marked with a black dot, represents the minimum sampling rate required for the successful reconstruction of a given value of rank R. The empirical probability of success for each point is 0.99. The empirical probability is computed over 100 iterations with a new instance of randomly generated $\boldsymbol{X}_0$ in each iteration. The red line shows the least-squares fit of the black points. It is clear from the plot that the for reasonably large values of R, the sampling rate is within a small constant of the optimal rate $R(W+M-R)$. In context of the application, and under the narrow-band assumption described in Section 1.2, the graph in Figure 9(b) shows that for a fixed number of sources $R = 10$, the sufficient sampling rate Ω required for the successful reconstruction of the ensemble decreases inversely with increasing number M of the receiving antennas. Each black point gives the minimum sampling rate required for the successful reconstruction with probability 0.99. The red line is the least-squares fit of these marked points. In other words, Figure 9(b) illustrates the relationship between the number of ADCs, or receiving antennas M, and the sampling rate Ω of each of the ADC for a fixed number of sources $R = 10$. The important point is that as we increase the number of antennas the the sampling burden on each of the ADCs decreases.



(a) (b)

Figure 9: Performance of the random demodulator for multiple signals lying in a subspace. In these experiments, we take an ensemble of 100 signals, each bandlimited to 512Hz. The probability of success is computed over 100 iterations. (a) Oversampling factor η as a function of the number R of underlying independent signals. The blue line is the least-squares fit of the data points. (b) Sampling rate Ω versus the number M of recieving antennas. The blue line is the least-sqaures fit of the data points.

3.2 Stable recovery

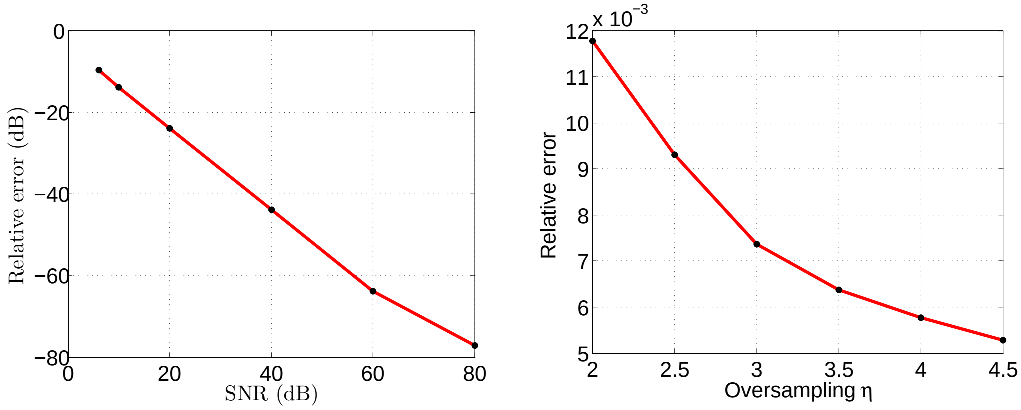
In the second set of experiments, we study the performance of the the recovery algorithm when the measurements are contaminated with noise as in (12). The noise vector is standard Gaussian, i.e., $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. We select $\delta \leq \sigma(L + \sqrt{L})^{1/2}$; a natural choice as the condition $\|\boldsymbol{\xi}\|_2 \leq \delta$ holds with high probability. In the first set of experiments shown in Figure 10, we solve the optimization program in (13). The plot in Figure 10(a) shows the relationship between the signal-to-noise ratio (SNR):

$$\text{SNR(dB)} = 10 \log \left(\frac{\|\boldsymbol{X}_0\|_{\text{F}}^2}{\|\boldsymbol{\xi}\|_2^2} \right),$$

and the realtive error(dB):

$$\text{Relative error (dB)} = 10 \log \left(\frac{\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\text{F}}^2}{\|\boldsymbol{X}_0\|_{\text{F}}^2} \right)$$

for a fixed oversampling factor $\eta = 3.5$. The result shows that the relative error degrades gracefully with decreasing SNR. In the Figure 10(b), the plot depicts relative error as a function of the oversampling factor for a fixed SNR = 40dB. The relative error decrease with increasing sampling rate.



(a) (b)

Figure 10: Recovery using matix Lasso in the presence of noise. The input ensemble to the simulated random demodulator consists of 100 signals, each bandlimited to 512Hz with number $R = 15$ of latent independent signals.(a) The SNR in dB versus the relative error in dB. The oversampling factor $\eta = 3.5$. (b) Relative error as a function of the sampling rate. The SNR is fixed at 40dB.

4 Proof of Lemma 1

We start with the proof of Lemma 1

Proof. The point (1) is the standard result [43]. We give proof of (2) now. It is a fact that in (29) and (30) for fixed a and $\theta \sim \text{Uniform}([0, 2\pi])$, the random variables $\text{sign}(\cos(a + \theta))$ and $\text{sign}(\sin(a + \theta))$ are independent of one another. Thus $\boldsymbol{H}$ has the same probability distribution as $\boldsymbol{W}\boldsymbol{Z}\boldsymbol{Q}^*$, where $\boldsymbol{Z} = \text{diag}(\boldsymbol{z})$ and the entries of $\mathbf{z}$ are i.i.d ± 1 random variables. In light of this, we will replace $\boldsymbol{H}$ with $\boldsymbol{W}\mathbf{Z}\boldsymbol{Q}^*$. For a fixed k, we can write

$$\tilde{\boldsymbol{V}}^* \boldsymbol{e}_k = \boldsymbol{V}^* \boldsymbol{H}^* \boldsymbol{e}_k = \hat{\boldsymbol{V}}^* \boldsymbol{Z} \boldsymbol{w}_k$$
$$= \sum_{\omega=1}^{W} \boldsymbol{z}(\omega) \boldsymbol{w}_k(\omega) \hat{\boldsymbol{v}}_\omega$$

where $\hat{\boldsymbol{V}} = \boldsymbol{Q}^* \boldsymbol{V}$ and $\boldsymbol{w}_k = \boldsymbol{W}^* \boldsymbol{e}_k$ and $\hat{\boldsymbol{v}}_\omega = \hat{\boldsymbol{V}}^* \boldsymbol{e}_\omega$ is the ωth row of $\hat{\boldsymbol{V}}$. We will apply the following concentration inequality,

Theorem 3. *[44] Let $\eta \in \mathbb{R}^n$ be a vector whose entries are independent random variables with $|\boldsymbol{\eta}(i)| < 1$, and let $\boldsymbol{S}$ be a fixed $m \times n$ matrix. Then for every $t \geq 0$*

$$\mathrm{P}\left\{ \|\boldsymbol{S}\boldsymbol{\eta}\|_2 \geq \mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\eta}\|_2 + t \right\} \leq 2e^{-t^2/16\|\boldsymbol{S}\|^2},$$

where

$$\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\eta}\|_2 \leq \|\boldsymbol{S}\|_{\mathrm{F}}\,.$$

We can apply the above theorem with $\boldsymbol{S} = \hat{\boldsymbol{V}}^* \boldsymbol{W}_k$, where $\boldsymbol{W}_k = \text{diag}(\boldsymbol{w}_k)$, and $\boldsymbol{\eta} = \boldsymbol{z}$. In this case, we have

$$\left\| \hat{\boldsymbol{V}}^* \boldsymbol{W}_k \right\|_{\mathrm{F}}^2 = \sum_{\omega=1}^{W} |\boldsymbol{w}_k(\omega)|^2 \, \|\hat{\boldsymbol{v}}_\omega\|_2^2$$
$$\leq \frac{2}{W} \sum_{\omega=1}^{W} \|\hat{\boldsymbol{v}}_\omega\|_2^2$$

$$\leq \frac{2R}{W},$$

and

$$\|\hat{\boldsymbol{V}}^* \boldsymbol{W}_k\| \leq \sqrt{\frac{2}{W}} \|\hat{\boldsymbol{V}}^*\| = \sqrt{\frac{2}{W}}.$$

Thus,

$$\mathrm{P}\left\{ \|\tilde{\boldsymbol{V}}^* \boldsymbol{e}_k\|_2 > \sqrt{\frac{2R}{W}} + t\sqrt{\frac{2}{W}} \right\} \leq 2e^{-t^2/16},$$

and

$$\mathrm{P}\left\{ \max_{1\leq k\leq W} \|\tilde{\boldsymbol{V}}^* \boldsymbol{e}_k\|_2 > \sqrt{\frac{2R}{W}} + t\sqrt{\frac{2}{W}} \right\} \leq 2We^{-t^2/16}.$$

We can make this probability less than $W^{-\beta}$ by taking $t \geq C\sqrt{\log W}$, and (2) follows.

Now for (3), we can write $\boldsymbol{H} = \boldsymbol{W}\boldsymbol{Z}\boldsymbol{Q}^*$. Let $\boldsymbol{w}_\ell$ be the ℓth column of $\boldsymbol{W}^*$ and let $\tilde{\boldsymbol{u}}_k^*$ be the kth row of $\tilde{\boldsymbol{U}}$. For a fixed row index k and column index ℓ, we can write an entry of $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^*$ as

$$\begin{aligned}\left(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^*\right)(k,\ell) &= \tilde{\boldsymbol{U}}(\boldsymbol{W}\boldsymbol{Z}\boldsymbol{Q}^*\boldsymbol{V})^* \\ &= \tilde{\boldsymbol{U}}\tilde{\boldsymbol{Q}}^*\boldsymbol{Z}\boldsymbol{W}^* \\ &= (\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{U}}_k)^*\boldsymbol{Z}\boldsymbol{w}_\ell \\ &= \sum_{\omega=1}^{W} (\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{u}}_k(\omega))^*\boldsymbol{z}(\omega)\boldsymbol{w}_\ell(\omega),\end{aligned}$$

where $\tilde{\boldsymbol{Q}} = \boldsymbol{Q}^*\boldsymbol{V}$ is a tall orthonormal matrix. Since the $\boldsymbol{z}(\omega)$ are i.i.d. random variables, a standard applications of the Hoeffding inequality tells us that

$$\mathrm{P}\left\{ \left| \left(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^*\right)(k,\ell) \right| > \lambda \right\} \leq 2e^{-\lambda^2/2\sigma^2},$$

where

$$\begin{aligned}\sigma^2 &= \sum_{\omega=1}^{W} \left| \left(\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{u}}_k(\omega)\right) \right|^2 |\boldsymbol{w}_\ell(\omega)|^2 \\ &\leq \frac{2\left\|\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{u}}_k\right\|_2^2}{W} \\ &= \frac{2\|\tilde{\boldsymbol{u}}_k\|_2^2}{W}.\end{aligned}$$

Thus, with probability exceeding $1 - 2W^{-\beta}$

$$\max_{\substack{1\leq k\leq M \\ 1\leq \ell\leq W}} \left| \left(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^*\right)(k,\ell) \right|^2 \leq \frac{4(\beta+2)\log W}{W} \max_{1\leq \ell\leq M} \|\tilde{\boldsymbol{u}}_k\|_2^2.$$

The point (1) tells us that

$$\max_{1\leq k\leq M} \|\tilde{\boldsymbol{u}}_k\|_2^2 \leq C_\beta \frac{\max(R, \log M)}{M}$$

with probability exceeding $1 - M^{-\beta}$. Thus (3) holds with probability exceeding $1 - O(W^{-\beta} + M^{-\beta})$. □

5 Proof of Theorem 1

Define a subspace $T \subset \mathbb{R}^{M\times W}$ associated with $\boldsymbol{X}_0$ with svd $\sum_{k=1}^{R}\sigma_k \boldsymbol{u}_k\boldsymbol{v}_k^*$:

$$T = \{\boldsymbol{X} : \boldsymbol{X} = \boldsymbol{U}\boldsymbol{Z}_1^* + \boldsymbol{Z}_2\boldsymbol{V}^*, \boldsymbol{Z}_1 \in \mathbb{R}^{W\times R}, \boldsymbol{Z}_2 \in \mathbb{R}^{M\times R}\}.$$

The orthogonal projection of $\mathcal{P}_T$ onto T is

$$\mathcal{P}_T(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{U}^*\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^* - \boldsymbol{U}\boldsymbol{U}^*\boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^*,$$

and its orthogonal complement

$$\mathcal{P}_{T^\perp}(\boldsymbol{Z}) = (\mathcal{I} - \mathcal{P}_T)(\boldsymbol{X}) = (\boldsymbol{I}_M - \boldsymbol{U}\boldsymbol{U}^*)\boldsymbol{X}(\boldsymbol{I}_W - \boldsymbol{V}\boldsymbol{V}^*).$$

A sufficient condition for the uniqueness of the minimizer to (11) is given by the following Proposition [3,4].

Proposition 1. *The matrix $\boldsymbol{X}$ is the unique minimizer to (11) if $\exists \boldsymbol{Y} \in$ Range$(\mathcal{A}^*)$ such that*

$$\|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* - \|\boldsymbol{U}\boldsymbol{V}^* - \mathcal{P}_T(\boldsymbol{Y})\|_{\mathrm{F}}\|\mathcal{P}_T(\boldsymbol{Z})\|_{\mathrm{F}} - \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\|\|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* > 0,$$

for all $\boldsymbol{Z} \in$ Null$(\mathcal{A})$.

Proposition 1 implies that the uniqueness of the minimizer is gauranteed, if $\exists \boldsymbol{Y} \in \mathrm{Range}(\mathcal{A}^*)$, such that

$$\|\mathcal{P}_T(\boldsymbol{Y}) - \boldsymbol{U}\boldsymbol{V}^*\|_{\mathrm{F}} \leq \sqrt{\frac{\Omega}{9W}}, \quad \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\| \leq \frac{1}{2}, \tag{33}$$

holds. Also for $\boldsymbol{Z} \neq \boldsymbol{0}$, and $\forall \boldsymbol{Z} \in \mathrm{Null}(\mathcal{A})$ the following

$$\|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_{\mathrm{F}} \geq \sqrt{\frac{\Omega}{2W}}\|\mathcal{P}_T(\boldsymbol{Z})\|_{\mathrm{F}} \tag{34}$$

is true. To show (34), for $\boldsymbol{Z} \in \mathrm{Null}(\mathcal{A})$

$$\begin{aligned} 0 &= \|\mathcal{A}(\boldsymbol{Z})\|_{\mathrm{F}} \\ 0 &\geq \|\mathcal{A}(\mathcal{P}_T(\boldsymbol{Z}))\|_{\mathrm{F}} - \|\mathcal{A}(\mathcal{P}_{T^\perp}(\boldsymbol{Z}))\|_{\mathrm{F}}, \end{aligned}$$

which after using the fact that $\|\mathcal{A}\| = \sqrt{W/\Omega}$ implies that

$$\|\mathcal{A}(\mathcal{P}_{T^\perp}(\boldsymbol{Z}))\|_{\mathrm{F}} \leq \sqrt{\frac{W}{\Omega}}\|\mathcal{P}_T^\perp(\boldsymbol{Z})\|_{\mathrm{F}}. \tag{35}$$

In addition, for an arbitrary $\boldsymbol{Z}$, we have

$$\begin{aligned} \|\mathcal{A}(\mathcal{P}_T(\boldsymbol{Z}))\|_{\mathrm{F}}^2 &= \langle \mathcal{A}(\mathcal{P}_T(\boldsymbol{Z})), \mathcal{A}(\mathcal{P}_T(\boldsymbol{Z}))\rangle \\ &= \langle \boldsymbol{Z}, \mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T(\boldsymbol{Z})\rangle \\ &\geq (1 - \|\mathcal{P}_T\mathcal{A}\mathcal{A}^*\mathcal{P}_T - \mathcal{P}_T\|)\|\mathcal{P}_T(\boldsymbol{Z})\|_{\mathrm{F}}^2 \\ &\geq \frac{1}{2}\|\mathcal{P}_T(\boldsymbol{Z})\|_{\mathrm{F}}^2, \end{aligned} \tag{36}$$

where the last inequality is obtained by plugging in $\|\mathcal{P}_T\mathcal{A}\mathcal{A}^*\mathcal{P}_T - \mathcal{P}_T\| \leq \frac{1}{2}$, which is true with probability at least $1 - O(WM)^{-\beta}$ by the application of Corollary 2, using $\Omega \geq C\beta R(\mu_1^2(W/M) + \mu_2^2)\log^2(WM)$. Collecting the facts in (35), and (36), the result in (34) is obtained.

5.1 Measurements as a matrix trace inner product

The (i,j)th sample taken by the ADC in the i-th branch in Figure 6 will be expressed using trace inner product as

$$y_{ij} = \langle \boldsymbol{A}_{ij}, \boldsymbol{X}_0 \rangle = \operatorname{tr}(\boldsymbol{A}_{ij}^* \boldsymbol{X}_0) = \sum_{k \in \mathcal{B}_j} d_i[k] X_0[i,k], \quad (i,j) = \{1,\dots,M\} \times \{1,\dots,\Omega\}, \tag{37}$$

where the sampling mask $\boldsymbol{A}_{ij}$ is

$$\boldsymbol{A}_{ij} = \sum_{k \sim \mathcal{B}_j} d_i[k] \boldsymbol{e}_i \tilde{\boldsymbol{e}}_k^*. \tag{38}$$

where $\{\boldsymbol{e}_i\}_{1 \le i \le M}$, and $\{\tilde{\boldsymbol{e}}_k\}_{1 \le k \le W}$ are standard basis vectors of dimension M, and W, respectively. It follows that

$$\mathcal{A}^*\mathcal{A}(\boldsymbol{X}) = \sum_{(i,j)} \langle \boldsymbol{A}_{ij}, \boldsymbol{X} \rangle \boldsymbol{A}_{ij},$$

and

$$\mathcal{A}^*\mathcal{A} = \sum_{(i,j)} \boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij}, \tag{39}$$

where $\otimes$ denotes the tensor product. It is clear that the measurement matrices $\boldsymbol{A}_{ij}$ are rank-1 random matrices. In general, the tensor product of rank-1 matrices $\boldsymbol{u}_1\boldsymbol{v}_1^*$, $\boldsymbol{u}_2\boldsymbol{v}_2^*$ with $\boldsymbol{u}_i \in \mathbb{R}^M$, and $\boldsymbol{v}_i \in \mathbb{R}^W$ is given by the big matrix

$$\boldsymbol{u}_1\boldsymbol{v}_1^* \otimes \boldsymbol{u}_2\boldsymbol{v}_2^* = \begin{bmatrix} u_1[1]^* u_2[1] \boldsymbol{v}_1\boldsymbol{v}_2^* & u_1[1]^* u_2[2] \boldsymbol{v}_1\boldsymbol{v}_2^* & \cdots & u_1[1]^* u_2[N] \boldsymbol{v}_1\boldsymbol{v}_2^* \\ u_1[2]^* u_2[1] \boldsymbol{v}_1\boldsymbol{v}_2^* & u_1[2]^* u_2[2] \boldsymbol{v}_1\boldsymbol{v}_2^* & \cdots & u_1[2]^* u_2[N] \boldsymbol{v}_1\boldsymbol{v}_2^* \\ \vdots & & \ddots & \\ u_1[N]^* u_2[1] \boldsymbol{v}_1\boldsymbol{v}_2^* & \cdots & \cdots & u_1[N]^* u_2[N] \boldsymbol{v}_1\boldsymbol{v}_2^*, \end{bmatrix}$$

and we will denote (α,β)th, $W \times W$ submatrix by

$$\{\boldsymbol{u}_1\boldsymbol{v}_1^* \otimes \boldsymbol{u}_2\boldsymbol{v}_2^*\}_{(\alpha,\beta)} = u_1[\alpha]^* u_2[\beta] \boldsymbol{v}_1\boldsymbol{v}_2^*.$$

Using the above notation, we can write

$$\{\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij}\}_{(\alpha,\beta)} = e_i[\alpha] e_i[\beta] \sum_{k,k' \sim \mathcal{B}_j} d_i[k] d_i[k'] \tilde{\boldsymbol{e}}_k \tilde{\boldsymbol{e}}_{k'}^*. \tag{40}$$

Taking the expectation, we can see that

$$\{\operatorname{E}(\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij})\}_{(\alpha,\beta)} = e_i[\alpha] e_i[\beta] \sum_{k \sim \mathcal{B}_j} \tilde{\boldsymbol{e}}_k \tilde{\boldsymbol{e}}_k^*,$$

and using the fact that $e_i[\alpha]e_i[\beta] = 1$ when $\alpha = \beta$ and is zero otherwise, we can see that

$$\sum_{(i,j)} \operatorname{E}(\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij}) = \boldsymbol{I}_{WM},$$

where $\boldsymbol{I}_{WM}$ denotes $WM \times WM$ identity matrix. In operator notation, we have $\operatorname{E}\mathcal{A}^*\mathcal{A} = \mathcal{I}$.

Let $\{\boldsymbol{u}_k^*\}_{1 \le k \le M}$, and $\{\boldsymbol{v}_k^*\}_{1 \le k \le W}$ denote the rows of the matrices $\boldsymbol{U}$, and $\boldsymbol{V}$, respectively. The following quantity will be used repeatedly in the theoretical analysis

$$\begin{aligned} \|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2 &= \langle \mathcal{P}_T(\boldsymbol{A}_{ij}), \boldsymbol{A}_{ij} \rangle \\ &= \langle \boldsymbol{U}^*\boldsymbol{A}_{ij}, \boldsymbol{U}^*\boldsymbol{A}_{ij} \rangle + \langle \boldsymbol{A}_{ij}\boldsymbol{V}, \boldsymbol{A}_{ij}\boldsymbol{V} \rangle - \langle \boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}, \boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V} \rangle \\ &= \|\boldsymbol{U}^*\boldsymbol{A}_{ij}\|_{\mathrm{F}}^2 + \|\boldsymbol{A}_{ij}\boldsymbol{V}\|_{\mathrm{F}}^2 - \|\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\|_{\mathrm{F}}^2 \le \|\boldsymbol{U}^*\boldsymbol{A}_{ij}\|_{\mathrm{F}}^2 + \|\boldsymbol{A}_{ij}\boldsymbol{V}\|_{\mathrm{F}}^2. \end{aligned}$$

Using the definition (38), we have

$$\|\boldsymbol{U}^*\boldsymbol{A}_{ij}\|_{\mathrm{F}}^2 = \left\|\sum_{k\sim\mathcal{B}_j} d_i[k]\boldsymbol{u}_i\tilde{\boldsymbol{e}}_k^*\right\|_{\mathrm{F}}^2 = \|\boldsymbol{u}_i\|_2^2\left\|\sum_{k\sim\mathcal{B}_j} d_i[k]\tilde{\boldsymbol{e}}_k\right\|_2^2 \le \mu_1^2\frac{R}{M}\cdot\frac{W}{\Omega},$$

and

$$\|\boldsymbol{A}_{ij}\boldsymbol{V}\|_{\mathrm{F}}^2 = \left\|\sum_{k\sim\mathcal{B}_j} d_i[k]\boldsymbol{e}_i\boldsymbol{v}_k^*\right\|_{\mathrm{F}}^2 = \|\boldsymbol{e}_i\|_2^2\left\|\sum_{k\sim\mathcal{B}_j} d_i[k]\boldsymbol{v}_k\right\|_2^2.$$

This implies that

$$\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2 \le \mu_1^2\frac{R(W/M)}{\Omega} + \left\|\sum_{k\sim\mathcal{B}_j} d_i[k]\boldsymbol{v}_k\right\|_2^2. \tag{41}$$

5.2 Golfing scheme for the random modulator

We start with partitioning the measurements indexed by the set

$$\Gamma = \{(i,j)\}_{\substack{1\le i\le M\\ 1\le j\le\Omega}}$$

into κ disjoint partitions $\{\Gamma_k\}_{1\le k\le\kappa}$ of size $|\Gamma_k| = L/\kappa$, such that $\bigcup_k \Gamma_k = \Gamma$, i.e., $\kappa|\Gamma_k| = M\Omega$. We will construct the dual certificate $\boldsymbol{Y} \in \mathrm{Range}(\mathcal{A}^*)$ iteratively using Gross's golfing scheme. Let $\mathcal{A}_k$ denote the operator corresponding to the samples taken in the kth partition, i.e.,

$$\mathcal{A}_k^*\mathcal{A}_k = \sum_{(i,j)\in\Gamma_k} \boldsymbol{A}_{ij}\otimes\boldsymbol{A}_{ij}.$$

As will be clear later in the proof that we want the partitioned linear operator $\kappa\mathcal{A}_k^*\mathcal{A}_k$ to be a close approximation of the $\mathcal{I}$. For this purpose, each of the partition Γ_k is chosen uniformly at random out of the set Γ. Suppose now that we form a new set of partitions $\{\Gamma_k'\}$ defined as

$$\Gamma_k' = \{(i,j)\in\{1,\dots,M\}\times\{1,\dots,\Omega\} : \delta_{(i,j)} = 1\}, \tag{42}$$

where the sequence $\{\delta_{(i,j)}\}_{\substack{1\le i\le M\\ 1\le j\le\Omega}}$ are independent 0/1 Bernoulli random variables with

$$\mathrm{P}\left\{\delta_{(i,j)} = 1\right\} = \frac{1}{\kappa}.$$

In the the proofs later, we will be interested in bounding events $\eta(\Gamma_k)$ that involve sum of independent random matrices indexed by the partitions $\{\Gamma_k\}_{1\le k\le\kappa}$, e.g., define

$$\eta(\Gamma_k) := \left\|\sum_{(i,j)\in\Gamma_k}\kappa\boldsymbol{A}_{ij}\otimes\boldsymbol{A}_{ij} - \mathcal{I}\right\|,$$

and we want to bound the probability $\mathrm{P}\{\eta(\Gamma_k) > \epsilon\}$. Uisng the fact that

$$\mathrm{P}\{\eta(\Gamma_k) > \epsilon\} \le 2\,\mathrm{P}\{\eta(\Gamma_k') > \epsilon\}, \tag{43}$$

which implies that probability of an event $\{\eta(\Gamma_k) > \epsilon\}$ over the set Γ_k can be bounded by the probability of a similar event $\{\eta(\Gamma_k') > \epsilon\}$ over the set Γ_k'. As a result, we will now be concerned with only bounding the probability of events of interest over the sets Γ_k'. Thus, we redefine $\mathcal{A}_k^*\mathcal{A}_k$ over Γ_k' as

$$\mathcal{A}_K^*\mathcal{A}_k = \sum_{(i,j)\in\Gamma_k'}\boldsymbol{A}_{ij}\otimes\boldsymbol{A}_{ij} = \sum_{(i,j)}\delta_{(i,j)}\boldsymbol{A}_{ij}\otimes\boldsymbol{A}_{ij}. \tag{44}$$

The iterative construction of the dual certificate is:

$$\boldsymbol{Y}_k = \boldsymbol{Y}_{k-1} - \kappa \mathcal{A}_k^* \mathcal{A}_k \left(\mathcal{P}_T(\boldsymbol{Y}_{k-1}) - \boldsymbol{U}\boldsymbol{V}^*\right).$$

Projecting on the subspace T on both sides results in

$$\mathcal{P}_T(\boldsymbol{Y}_k) = \mathcal{P}_T(\boldsymbol{Y}_{k-1}) - \kappa \mathcal{P}_T \mathcal{A}_k^* \mathcal{A}_k (\mathcal{P}_T(\boldsymbol{Y}_{k-1}) - \boldsymbol{U}\boldsymbol{V}^*),$$

where it is importatnt to see that $\boldsymbol{Y}_k \in \text{Range}(\mathcal{A}^*)$. Now let

$$\boldsymbol{W}_k := \mathcal{P}_T(\boldsymbol{Y}_k) - \boldsymbol{U}\boldsymbol{V}^*, \tag{45}$$

which gives

$$\boldsymbol{W}_k = \boldsymbol{W}_{k-1} - \kappa \mathcal{P}_T \mathcal{A}_k^* \mathcal{A}_k \mathcal{P}_T(\boldsymbol{W}_{k-1}).$$

As a result,

$$\|\boldsymbol{W}_k\|_{\text{F}} \le \|\kappa \mathcal{P}_T \mathcal{A}_k^* \mathcal{A}_k \mathcal{P}_T - \mathcal{P}_T\| \|\boldsymbol{W}_{k-1}\|_{\text{F}},$$

and by Lemma 2 with $\Omega \ge C\beta\kappa R(\mu_1^2(W/M) + \mu_2^2)\log^2(WM)$, it follows that

$$\begin{aligned} \|\boldsymbol{W}_\kappa\|_{\text{F}} &\le \left(\frac{1}{2}\right)^\kappa \|\boldsymbol{U}\boldsymbol{V}^*\|_{\text{F}} \\ &= 2^{-\kappa}\sqrt{R} \le \sqrt{\frac{\Omega}{9W}}, \quad \text{when} \quad \kappa \ge 0.5\log_2\left(\frac{9WR}{\Omega}\right), \end{aligned} \tag{46}$$

which holds with probability $1 - (WM)^{-\beta}$. In view of the coherences of $\boldsymbol{W}_0 = -\boldsymbol{U}\boldsymbol{V}^*$ with $\|\boldsymbol{U}\boldsymbol{V}^*\|_{\text{F}}^2 = R$ defined in (21), the coherence $\mu_{3,k}^2$ is related to the Frobenius norm of $\boldsymbol{W}_k$ as

$$\max_{i,j} \sum_{k\sim\mathcal{B}_j} \langle \boldsymbol{W}_k, \boldsymbol{e}_i \tilde{\boldsymbol{e}}_k^* \rangle^2 = \mu_{3,k}^2 \|\boldsymbol{W}_k\|_{\text{F}}^2 \frac{1}{M\Omega}. \tag{47}$$

Note that we have replaced R in the definition (15) with $\|\boldsymbol{W}_k\|_{\text{F}}^2$ for proper normalization. Lemma 4 shows that under appropriate conditions, the conclusion $\mu_{3,k}^2 \le \frac{1}{2}\mu_{3,k-1}^2$ holds with high probability. This implies that

$$\mu_{3,\kappa}^2 \le \mu_3^2 \tag{48}$$

is true and this fact will be used towards the end of this proof. The iterative dual certificate

$$\boldsymbol{Y} = \boldsymbol{Y}_\kappa = -\sum_{k=1}^{\kappa} \kappa \mathcal{A}_k^* \mathcal{A}_k (\boldsymbol{W}_{k-1})$$

satisfies (33). To show that $\|\mathcal{P}_{T^\perp}(\boldsymbol{Y}_\kappa)\| \le \frac{1}{2}$ holds given Lemma 3, and (48), we make the following calculation

$$\begin{aligned} \|\mathcal{P}_{T^\perp}(\boldsymbol{Y}_\kappa)\| &\le \sum_{k=1}^{\kappa} \|\mathcal{P}_{T^\perp}(\kappa \mathcal{A}_k^* \mathcal{A}_k(\boldsymbol{W}_{k-1}))\| = \sum_{k=1}^{\kappa} \|\mathcal{P}_{T^\perp}\left(\kappa \mathcal{A}_k^* \mathcal{A}_k(\boldsymbol{W}_{k-1}) - \boldsymbol{W}_{k-1}\right)\| \\ &\le \sum_{k=1}^{\kappa} \left(\kappa \mathcal{A}_k^* \mathcal{A}_k(\boldsymbol{W}_{k-1}) - \boldsymbol{W}_{k-1}\right)\| \le \sum_{k=1}^{\kappa} 2^{-k-1} < 1/2, \end{aligned}$$

which holds given $\Omega \ge C\beta\kappa R(W/M)\mu_3^2 \max(W/M, 1)\log^2(WM)$, the result holds with probabiltiy at least $1 - (WM)^{-\beta}$.

5.3 Lemmas for Theorem 1

We state here the key Lemmas required to prove sampling Theorem 1.

Lemma 2. *Suppose Ω measurements are taken through the random demodulator using the setup in* (18)*. Let $\mathcal{A}_k^*\mathcal{A}_k$, defined in* (44)*, be the kth partition of $\mathcal{A}^*\mathcal{A}$ indexed by $\Gamma_{k'}$, defined in* (42)*. Then for all $\beta > 1$,*

$$\|\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathcal{P}_T\| \leq \frac{1}{2}$$

provided $\Omega \geq C\beta\kappa R(\mu_1^2(W/M) + \mu_2^2)\log^2(WM)$ with probability at least $1 - (WM)^{-\beta}$.

Corollary 2. *Suppose Ω measurements are taken through the random demodulator using the setup in* (18)*. Let $\mathcal{A}^*\mathcal{A}$ be as defined in* (39)*. Then for all $\beta > 1$,*

$$\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T - \mathcal{P}_T\| \leq \frac{1}{2}$$

provided $\Omega \geq C\beta R(\mu_1^2(W/M) + \mu_2^2)\log^2(WM)$ with probability at least $1 - (WM)^{-\beta}$.

Proof. Proof of the corollary follows from the proof of Lemma 2 without partitioning, i.e., $\kappa = 1$. □

Lemma 3. *Suppose Ω entries are observed using the random demodulator, as in* (18)*. Let $\boldsymbol{W}_{k-1}$ be a fixed $M \times W$ matrix defined in* (45)*. Then for all $\beta > 1$,*

$$\|(\kappa\mathcal{A}_k^*\mathcal{A}_k - \mathcal{I})(\boldsymbol{W}_{k-1})\| \leq 2^{-k-1}$$

with probability at least $1 - (WM)^{-\beta}$ provided $\Omega \geq C\beta\kappa\mu_{3,k-1}^2 R\max(W/M, 1)\log^{3/2}(WM)$.

Lemma 4. *Let $\mu_{3,k}^2$ be the coherence of the iterates as defined in* (47)*. Then*

$$\mu_{3,k}^2 \leq \frac{1}{2}\mu_{3,k-1}^2$$

holds when $\Omega \geq C\beta\kappa R(\mu_1^2(W/M) + \mu_2^2)\log(WM)$ for $\beta > 1$ with probability at least $1 - (WM)^{-\beta}$.

5.4 Matrix Bernstein-type inequality

We will use a specialized version of the matrix Bernstein-type inequality [41,45] to bound the operator norm of the random matrices in this paper. The version of Bernstein listed below depends on the Orlicz norms $\|\boldsymbol{Z}\|_{\psi_\alpha},\ \alpha \geq 1$ of a matrix $\boldsymbol{Z}$. The Orlicz norms are deined as

$$\|\boldsymbol{Z}\|_{\psi_\alpha} = \inf\{u > 0 : \operatorname{E}\exp(\frac{\|\boldsymbol{Z}\|^\alpha}{u^\alpha}) \leq 2\}, \quad \alpha \geq 1. \tag{49}$$

Suppose that, for some constant $U_\alpha > 0, \|\boldsymbol{Z}_q\|_{\psi_\alpha} \leq U_{(\alpha)}, q = 1, \ldots, Q$ then the following proposition holds.

Proposition 2. *Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_Q$ be iid random matrices with dimensions $M \times N$ that satisfy $\operatorname{E}(\boldsymbol{Z}_q) = 0$. Suppose that $\|\boldsymbol{Z}\|_{\psi_\alpha} < \infty$ for some $\alpha \geq 1$. Define*

$$\sigma_Z = \max\left\{\left\|\sum_{q=1}^{Q}(\operatorname{E}\boldsymbol{Z}_q\boldsymbol{Z}_q^*)\right\|^{1/2}, \left\|\sum_{q=1}^{Q}(\operatorname{E}\boldsymbol{Z}_q^*\boldsymbol{Z}_q)\right\|^{1/2}\right\} \tag{50}$$

Then $\exists$ a constant $C > 0$ such that , for all $t > 0$, with probability at least $1 - \mathrm{e}^{-t}$

$$\|\boldsymbol{Z}_1 + \cdots + \boldsymbol{Z}_Q\| \leq C\max\left\{\sigma_Z\sqrt{t + \log(M+N)}, U_\alpha\log^{1/\alpha}\left(\frac{QU_\alpha^2}{\sigma_Z^2}\right)(t + \log(M+N))\right\} \tag{51}$$

6 Proof of Lemmas for Theorem 1

6.1 Proof of Lemma 2

We want to bound the quantity

$$\eta(\Gamma_k) := \|\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathcal{P}_T\| = \left\|\sum_{(i,j)\in\Gamma_k}\kappa\mathcal{P}_T(\boldsymbol{A}_{ij})\otimes\mathcal{P}_T(\boldsymbol{A}_{ij}) - \mathcal{P}_T\right\|.$$

We are interested in the failure probability of the event $F(\Gamma_k) := \{\eta(\Gamma_k) > \zeta\}$. From (43), it is clear that

$$\mathrm{P}\{F(\Gamma_k)\} \le 2\,\mathrm{P}\{F(\Gamma_k')\},$$

where the set Γ_k', as defined in (42), is the partition generated using the Bernoulli model. Hence, it is enough to bound the operator norm

$$\left\|\sum_{(i,j)\in\Gamma_k'}\kappa\mathcal{P}_T(\boldsymbol{A}_{ij})\otimes\mathcal{P}_T(\boldsymbol{A}_{ij}) - \mathcal{P}_T\right\| = \left\|\sum_{(i,j)}\kappa\delta_{(i,j)}\mathcal{P}_T(\boldsymbol{A}_{ij})\otimes\mathcal{P}_T(\boldsymbol{A}_{ij}) - \mathcal{P}_T\right\|.$$

Now the fact that the operator $\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T$ does not deviate from its expected value

$$\begin{aligned}\mathrm{E}(\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T) &= \kappa\mathcal{P}_T\,\mathrm{E}\sum_{(i,j)\in\Gamma_k'}\boldsymbol{A}_{ij}\otimes\boldsymbol{A}_{ij}\mathcal{P}_T\\ &= \kappa\mathcal{P}_T\sum_{(i,j)}\mathrm{E}\,\delta_{(i,j)}\,\mathrm{E}(\boldsymbol{A}_{ij}\otimes\boldsymbol{A}_{ij})\mathcal{P}_T = \mathcal{P}_T\,\mathrm{E}\,\mathcal{A}^*\mathcal{A}\mathcal{P}_T = \mathcal{P}_T\end{aligned}$$

in the spectral norm can be proven using the matrix Bernstein Inequality. To proceed define the operator $\mathcal{L}_{ij}$ which maps $\boldsymbol{Z}$ to $\langle\mathcal{P}_T(\boldsymbol{A}_{ij}),\boldsymbol{Z}\rangle\,\mathcal{P}_T(\boldsymbol{A}_{ij})$, i.e., $\mathcal{L}_{ij} = \mathcal{P}_T(\boldsymbol{A}_{ij})\otimes\mathcal{P}_T(\boldsymbol{A}_{ij})$. This operator is rank one, therefore, the operator norm $\|\mathcal{L}_{ij}\| = \|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2$. Let

$$\boldsymbol{Z}_{ij} = \kappa\delta_{(i,j)}\mathcal{L}_{ij} - \kappa\,\mathrm{E}\,\delta_{(i,j)}\,\mathrm{E}\,\mathcal{L}_{ij}$$

We also have

$$\begin{aligned}\sum_{(i,j)}\kappa^2\,\mathrm{E}(\mathcal{L}_{ij} - \mathrm{E}\,\mathcal{L}_{ij})^2 &= \sum_{(i,j)}\kappa^2[\mathrm{E}(\delta_{(i,j)}\mathcal{L}_{ij}^2) - (\mathrm{E}\,\delta_{(i,j)}\mathcal{L}_{ij})^2]\\ &= \sum_{i,j}\kappa^2\,\mathrm{E}(\delta_{(i,j)}\,\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2\,\mathcal{L}_{ij}) - \sum_{i,j}\kappa^2(\mathrm{E}\,\delta_{(i,j)}\mathcal{L}_{ij})^2.\end{aligned}$$

Because $\mathrm{E}\,\mathcal{L}_{ij}^2$, and $(\mathrm{E}\,\mathcal{L}_{ij})^2$ are symmetric, positive-semidefinite matrices, it follows that

$$\begin{aligned}\left\|\sum_{i=1}^{M}\sum_{j=1}^{\Omega}\mathrm{E}\,\delta_{(i,j)}(\mathcal{L}_{ij} - \mathrm{E}\,\mathcal{L}_{ij})^2\right\| &\le \left\|\sum_{i,j}\mathrm{E}\,\delta_{(i,j)}\,\mathrm{E}\,\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2\,\mathcal{L}_{ij}\right\|\\ &= \frac{1}{\kappa}\left\|\sum_{i,j}\mathrm{E}\,\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2\,\mathcal{L}_{ij}\right\|\end{aligned}$$

Using the facts $\mathcal{L}_{ij} = \mathcal{P}_T(\boldsymbol{A}_{ij})\otimes\mathcal{P}_T(\boldsymbol{A}_{ij})$, $\sum_{ij}\mathrm{E}\,\mathcal{L}_{ij} = \mathcal{P}_T$, and expanding further gives

$$\left\|\mathrm{E}\sum_{i,j}\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2\,\mathcal{L}_{ij}\right\| \le \left\|\mathrm{E}\left(\sum_{i,j}\mu_1^2\frac{R(W/M)}{\Omega}\mathcal{L}_{ij} + \left\|\sum_{k\sim\mathcal{B}_j}d_i[k]\boldsymbol{v}_k\right\|_2^2\mathcal{L}_{ij}\right)\right\|.$$

Use the notation

$$\rho_{ij}^2 = \left\| \sum_{k\sim\mathcal{B}_j} d_i[k]\boldsymbol{v}_k \right\|_2^2,$$

then

$$\begin{aligned}\left\| \mathrm{E} \sum_{i,j} \|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2 \mathcal{L}_{ij} \right\| &\leq \mu_1^2 \frac{R(W/M)}{\Omega} \left\| \sum_{i,j} \mathrm{E}\,\mathcal{L}_{ij} \right\| + \left\| \mathrm{E} \sum_{i,j} \rho_{ij}^2 \mathcal{L}_{ij} \right\| \\ &\leq \mu_1^2 \frac{R(W/M)}{\Omega} + \left\| \mathrm{E} \sum_{i,j} \rho_{ij}^2 \mathcal{L}_{ij} \right\| \qquad (52)\end{aligned}$$

The second term in (52) can be simplified as

$$\begin{aligned}\left\| \mathrm{E} \sum_{i,j} \rho_{ij}^2 \mathcal{L}_{ij} \right\| &= \left\| \mathcal{P}_T\, \mathrm{E} \sum_{i,j} \left(\rho_{ij}^2 (\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij}) \right) \mathcal{P}_T \right\| \\ &\leq \left\| \mathrm{E} \sum_{i,j} \left(\rho_{ij}^2 (\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij}) \right) \right\|, \qquad (53)\end{aligned}$$

where the last line follows form the fact that $\|\mathcal{P}_T\| \leq 1$. Using the definition of ρ_{ij}^2, and (40), it is easy to see that the $W \times W$ submatrix at (α, β)-th location is given by

$$\{\mathrm{E}\left(\rho_{ij}^2(\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij})\right)\}_{(\alpha,\beta)} = e_i[\alpha]e_i[\beta] \left[\sum_{k\sim\mathcal{B}_j} \|\boldsymbol{v}_k\|_2^2 \sum_{\ell\sim\mathcal{B}_j} \tilde{\boldsymbol{e}}_\ell \tilde{\boldsymbol{e}}_\ell^* + \sum_{k\neq k'\sim\mathcal{B}_j} 2\langle \boldsymbol{v}_k, \boldsymbol{v}_{k'} \rangle \tilde{\boldsymbol{e}}_k \tilde{\boldsymbol{e}}_{k'}^* \right]. \qquad (54)$$

The following identity is very useful

$$\langle \boldsymbol{A}_{ij}, \boldsymbol{A}_{i'j'} \rangle = 0,$$

which holds true when either $i \neq i'$, or/and $j \neq j'$. Given this fact, we have

$$\| \sum_{i,j} \boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij} \| = \max_{ij} \|\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij}\|.$$

Using this fact, we can write

$$\left\| \sum_{i,j} \mathrm{E}\left(\rho_{ij}^2(\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij})\right) \right\| = \max_{i,j} \left\| \mathrm{E}\left(\rho_{ij}^2(\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij})\right) \right\|.$$

Using (54), we obtain

$$\begin{aligned}\left\| \sum_{i,j} \mathrm{E}\left(\rho_{ij}^2(\boldsymbol{A}_{ij} \otimes \boldsymbol{A}_{ij})\right) \right\| &\leq \left\| \sum_{k\sim\mathcal{B}_j} \|\boldsymbol{v}_k\|_2^2 \sum_{\ell\sim\mathcal{B}_j} \tilde{\boldsymbol{e}}_\ell \tilde{\boldsymbol{e}}_\ell^* \right\| + 2 \left\| \sum_{k\neq k'\sim\mathcal{B}_j} \langle \boldsymbol{v}_k, \boldsymbol{v}_{k'} \rangle \tilde{\boldsymbol{e}}_k \tilde{\boldsymbol{e}}_{k'}^* \right\| \\ &\leq \sum_{k\sim\mathcal{B}_j} \|\boldsymbol{v}_k\|_2^2 \left\| \sum_{\ell\sim\mathcal{B}_j} \tilde{\boldsymbol{e}}_\ell \tilde{\boldsymbol{e}}_\ell^* \right\| + 2\|\boldsymbol{v}_k\|_2^2 \left\| \sum_{k\neq k'\sim\mathcal{B}_j} \tilde{\boldsymbol{e}}_k \tilde{\boldsymbol{e}}_{k'}^* \right\| \\ &= \sum_{k\sim\mathcal{B}_j} \|\boldsymbol{v}_k\|_2^2 + 2\frac{W}{\Omega} \|\boldsymbol{v}_k\|_2^2 \leq 3\mu_2^2 \frac{R}{\Omega}, \qquad (55)\end{aligned}$$

where the second inequality follows from the application of Cauchy-Schwartz inequality in the second term of the R.H.S., the last equality is the result of the facts that

$$\left\|\sum_{\ell\sim\mathcal{B}_j}\tilde{\boldsymbol{e}}_\ell\tilde{\boldsymbol{e}}_\ell^*\right\| = 1, \quad \text{and} \quad \left\|\sum_{k\neq k'\sim\mathcal{B}_j}\tilde{\boldsymbol{e}}_k\tilde{\boldsymbol{e}}_{k'}^*\right\| \leq \frac{W}{\Omega}$$

and the last inequality follows form the definition of the coherence in (20). Plugging (55) in (52), we have the bound

$$\begin{aligned}\left\|\mathrm{E}\sum_{i,j}\boldsymbol{Z}_{ij}^*\boldsymbol{Z}_{ij}\right\| &\leq \kappa\left\|\mathrm{E}\sum_{i,j}\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2\,\mathcal{L}_{ij}\right\| \\ &\leq C\kappa\frac{R(\mu_1^2(W/M)+\mu_2^2)}{\Omega}. \end{aligned}\tag{56}$$

Finally, we calculate the Orlicz norm, the last ingredient to obtain the Bernstein bound. First, it is important to see that

$$\|\boldsymbol{Z}_{ij}\| \leq \kappa\|\mathcal{L}_{ij} - \mathrm{E}\,\mathcal{L}_{ij}\| \leq 2\kappa\|\mathcal{L}_{ij}\| = 2\|\mathcal{L}_{ij}\|_{\mathrm{F}} = 2\kappa\|\mathcal{P}_T(\boldsymbol{A}_{ij})\|_{\mathrm{F}}^2,$$

where the second-last equality follows form the fact that $\mathcal{L}_{ij}$ is the rank-1 operator. Using the last equation, and (41), we have

$$\begin{aligned}U_1 := \|\boldsymbol{Z}_{ij}\|_{\psi_1} &\leq 2\kappa\left\|\mu_1^2R\frac{(W/M)}{\Omega} + \sum_{r=1}^{R}\left(\sum_{\gamma\sim\mathcal{B}_j}d_i[k]v_{kr}\right)^2\right\|_{\psi_1} \\ &\leq C\mu_1^2\kappa R\frac{(W/M)}{\Omega} + C\kappa\sum_{k\sim\mathcal{B}_j}\|\boldsymbol{v}_k\|_2^2 \\ &\leq C\mu_1^2\kappa R\frac{(W/M)}{\Omega} + C\mu_2^2\kappa\frac{R}{\Omega}. \end{aligned}\tag{57}$$

Using the notation $\Lambda = \mu_1^2(W/M) + \mu_2^2$, we obtain

$$U_1\log\left(\frac{M\Omega\cdot U_1^2}{\sigma_Z^2}\right) = C\kappa R\frac{\Lambda}{\Omega}\log(\kappa RM\Lambda).$$

Plugging (56), and (57), and using $t = \beta\log(WM)$ in the non-commutative Bernstein's Inequality in Proposition 2, we have

$$\left\|\sum_{i=1}^{M}\sum_{j=1}^{\Omega}\boldsymbol{Z}_{ij}\right\| \leq 2\max\{\sqrt{\kappa R\frac{\Lambda}{\Omega}}\sqrt{\beta\log(WM)}, \kappa R\frac{\Lambda}{\Omega}\log(\kappa RM\Lambda)(\beta\log(WM))\}$$

The claim follwis by taking $t = \beta\log(WM)$, and the fact that $RM\Lambda \leq WM$, and $\Omega \geq C(\mu_1^2(W/M) + \mu_2^2)\kappa R\beta\log^2(WM)$ with probabiliy at least $1-(WM)^{-\beta}$.

6.2 Proof of Lemma 3

We will use the Bernstein bound in Proposition 2 to prove this Lemma. We want to bound

$$\|(\kappa\mathcal{A}_k^*\mathcal{A}_k - \mathcal{I})(\boldsymbol{W}_{k-1})\| = \left\|\sum_{(i,j)\in\Gamma_k}\kappa\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij} - \mathrm{E}\sum_{(i,j)\in\Gamma_k}\kappa\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij}\right\|,$$

which follows from

$$\boldsymbol{W}_{k-1} = \mathrm{E}\sum_{(i,j)\in\Gamma_k}\kappa\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij}.$$

Using the reasoning similar to that in Lemma 2, it is clear that bounding the following sum of random matrices over Γ_k' matrices

$$\left\|\sum_{(i,j)\in\Gamma_k'}\kappa\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij} - \mathrm{E}\sum_{(i,j)\in\Gamma_k}\kappa\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij}\right\|$$

suffices. Define a zero-mean random variable as follows:

$$\boldsymbol{Z}_{ij} = \kappa\delta_{(i,j)}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij} - \kappa\,\mathrm{E}\,\delta_{(i,j)}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij}.$$

The first variance term in (50) is

$$\begin{aligned}\sum_{(i,j)}\mathrm{E}\,\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^* = &\sum_{(i,j)}\kappa^2\,\mathrm{E}\,\delta_{(i,j)}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^*\\ &-\sum_{i,j}\kappa^2(\mathrm{E}\,\delta_{(i,j)})^2\,\mathrm{E}(\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij})\,\mathrm{E}(\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle\boldsymbol{A}_{ij})^*,\end{aligned}$$

where

$$\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle = \sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]W_{k-1}[i,\gamma].$$

The following can be easily verified and will be used in the proof of this Lemma

$$\begin{aligned}\left\|\sum_{(i,j)}\mathrm{E}\,\delta_{(i,j)}\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^*\right\| &\leq \left\|\sum_{(i,j)}\kappa^2\,\mathrm{E}\,\delta_{(i,j)}\,\mathrm{E}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^*\right\|\\ &= \left\|\sum_{(i,j)}\kappa\,\mathrm{E}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^*\right\|.\end{aligned}$$

In the calculations below, we assemble the ingredients to calculate the variance. First by (38), we have

$$\begin{aligned}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^* &= \langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{e}_i\boldsymbol{e}_i^*\sum_{\gamma,\gamma'\sim\mathcal{B}_j}d_i[k]d_i[k']\tilde{\boldsymbol{e}}_k^*\tilde{\boldsymbol{e}}_{k'}\\ &= \langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{e}_i\boldsymbol{e}_i^*\sum_{\gamma\sim\mathcal{B}_j}d_i[k]^2\|\tilde{\boldsymbol{e}}_k\|_2^2\\ &= \frac{W}{\Omega}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{e}_i\boldsymbol{e}_i^*.\end{aligned}$$

Taking summation over i, and j on both sides and using above relation gives us

$$\begin{aligned}\frac{W}{\Omega}\left\|\sum_{(i,j)}\mathrm{E}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\boldsymbol{e}_i\boldsymbol{e}_i^*\right\| &\leq \left\|\sum_{i=1}^M\boldsymbol{e}_i\boldsymbol{e}_i^*\right\|\cdot\max_i\frac{W}{\Omega}\sum_{j=1}^{\Omega}\mathrm{E}\langle\boldsymbol{A}_{ij},\boldsymbol{W}_{k-1}\rangle^2\\ &= \frac{W}{\Omega}\max_i\sum_{j=1}^{\Omega}\sum_{\gamma\sim\mathcal{B}_j}W_{k-1}^2[i,\gamma].\end{aligned}$$

This gives the first term in the variance (50)

$$\left\|\sum_{(i,j)}\mathrm{E}\,\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^*\right\| \leq \kappa\frac{W}{\Omega}\sum_{j=1}^{\Omega}\sum_{\gamma\sim\mathcal{B}_j}W_{k-1}^2[i,\gamma] \leq \|\boldsymbol{W}_{k-1}\|_{\mathrm{F}}^2\kappa\frac{(W/M)}{\Omega}\mu_{3,k-1}^2, \tag{58}$$

where the last inequality follows from (19). The second variance term in (50) is

$$\begin{aligned}\sum_{(i,j)} \mathrm{E}\, \boldsymbol{Z}_{ij}^* \boldsymbol{Z}_{ij} &= \sum_{(i,j)} \kappa^2 \,\mathrm{E}\, \delta_{(i,j)} \,\mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \boldsymbol{A}_{ij}^* \boldsymbol{A}_{ij} - \\ &- \sum_{i,j} \mathrm{E}\, \delta_{(i,j)} \,\mathrm{E}(\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij})^* \,\mathrm{E}(\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}) \\ &= \kappa \sum_{i,j} \mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \boldsymbol{A}_{ij}^* \boldsymbol{A}_{ij} - \kappa \sum_{i,j} \mathrm{E}(\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij})^* \,\mathrm{E}(\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}),\end{aligned}$$

which implies that

$$\left\| \sum_{(i,j)} \mathrm{E}\, \boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^* \right\| \leq \kappa \left\| \sum_{(i,j)} \mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \boldsymbol{A}_{ij} \boldsymbol{A}_{ij}^* \right\|.$$

We begin with

$$\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \boldsymbol{A}_{ij}^* \boldsymbol{A}_{ij} = \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \|\boldsymbol{e}_i\|_2^2 \sum_{\gamma,\gamma' \sim \mathcal{B}_j} d_i[\gamma] d_i[\gamma'] \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_{\gamma'}^*.$$

The expectation of the operand on the right hand side gives

$$\begin{aligned}\mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \boldsymbol{A}_{ij}^* \boldsymbol{A}_{ij} &= \left(\sum_\gamma d_i[\gamma] W_{k-1}[i,\gamma]\right)^2 \left(\sum_{\gamma,\gamma' \sim \mathcal{B}_j} d_i[\gamma] d_i[\gamma'] \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_{\gamma'}^*\right) \\ &= \sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma] \sum_{\gamma \sim \mathcal{B}_j} \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_\gamma^* + 2 \sum_{\gamma,\gamma' \sim \mathcal{B}_j} W_{k-1}[i,\gamma] W_{k-1}[i,\gamma'] \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_{\gamma'}^*.\end{aligned}$$

Next step is to take the operator norm, and summation over j on both sides. The orthogonality of $\{\tilde{\boldsymbol{e}}_\gamma\}_{1 \leq \gamma \leq W}$, and $\mathcal{B}_j \cap \mathcal{B}_{j'} = \phi$ implies that

$$\begin{aligned}&\left\| \sum_{j=1}^{\Omega} \mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \boldsymbol{A}_{ij}^* \boldsymbol{A}_{ij} \right\| \leq \\ &\leq \left\| \sum_{j=1}^{\Omega} \sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma] \sum_{\gamma \sim \mathcal{B}_j} \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_\gamma^* \right\| + 2 \left\| \sum_{j=1}^{\Omega} \sum_{\gamma,\gamma' \sim \mathcal{B}_j} W_{k-1}[i,\gamma] W_{k-1}[i,\gamma'] \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_{\gamma'}^* \right\| \\ &\leq \max_j \left\| \sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma] \sum_{\gamma \sim \mathcal{B}_j} \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_\gamma^* \right\| + 2 \max_j \left\| \sum_{\gamma,\gamma' \sim \mathcal{B}_j} W_{k-1}[i,\gamma] W_{k-1}[i,\gamma'] \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_{\gamma'}^* \right\|\end{aligned}$$

Now using the fact that

$$\left\| \sum_{\gamma,\gamma' \sim \mathcal{B}_j} W_{k-1}[i,\gamma] W_{k-1}[i,\gamma'] \tilde{\boldsymbol{e}}_\gamma \tilde{\boldsymbol{e}}_{\gamma'}^* \right\| = \sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma]$$

Summing over i, we obtain the second variance

$$\begin{aligned}\left\| \sum_{(i,j)} \mathrm{E}\, \boldsymbol{Z}_{ij}^* \boldsymbol{Z}_{ij} \right\| &\leq 3\kappa \sum_{i=1}^{M} \max_j \sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma] \\ &\leq 3\kappa \sum_{i=1}^{M} \|\boldsymbol{W}_{k-1}\|_{\mathrm{F}}^2 \mu_{3,k-1}^2 \frac{1}{M\Omega} = 3\kappa \mu_{3,k-1}^2 \frac{1}{\Omega} \|\boldsymbol{W}_{k-1}\|_{\mathrm{F}}^2. \qquad (59)\end{aligned}$$

Given (58), and (59), we can obtain the variance using (50). We now calculate the ψ_2 norm of matrix $\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij} - \mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}$. Since $\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}$ is a rank one matrix, this gives

$$\begin{aligned}\|\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij} - \mathrm{E}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}\|_{\psi_2} &\leq 2\left\|\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}\right\|_{\psi_2}\\ &\leq 2\|\boldsymbol{A}_{ij}\|\|\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle\|_{\psi_2}\\ &= 2\frac{W}{\Omega}\|\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle\|_{\psi_2}.\end{aligned}$$

This gives the result

$$\begin{aligned}U_2^2 &= \max_{ij} \kappa^2 \|\delta_{(i,j)}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij} - \mathrm{E}\,\delta_{(i,j)}\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle \boldsymbol{A}_{ij}\|_{\psi_2}^2\\ &\leq C\kappa \frac{W}{\Omega}\max_{ij}\sum_{\gamma\sim\mathcal{B}_j} W_{k-1}^2[i,\gamma] = C\kappa\frac{(W/M)}{\Omega^2}\mu_{3,k-1}^2\|\boldsymbol{W}_{k-1}\|_{\mathrm{F}}^2,\end{aligned} \tag{60}$$

and hence

$$U_2 \log^{1/2}\left(\frac{M\Omega\cdot U_2^2}{\sigma_Z^2}\right) \leq \sqrt{C\kappa\mu_{3,k-1}^2\frac{W/M}{\Omega^2}}\|\boldsymbol{W}_{k-1}\|_{\mathrm{F}}\log^{1/2}(WM).$$

The results in (58), (59), and (60) can be plugged in Proposition 2 to obtain

$$\begin{aligned}&\|\mathcal{A}_k^*\mathcal{A}_k(\boldsymbol{W}_{k-1}) - \boldsymbol{W}_{k-1}\|_{\mathrm{F}} \leq\\ &C\kappa\|\boldsymbol{W}_{k-1}\|_{\mathrm{F}}\max\left\{\sqrt{\frac{\kappa\mu_{3,k-1}^2\max(W/M,1)}{\Omega}}\sqrt{\beta\log(WM)}, \sqrt{\kappa\frac{\mu_{3,k-1}^2(W/M)}{\Omega^2}}\beta\log^{3/2}(WM)\right\}\end{aligned} \tag{61}$$

with $t = \beta\log(WM)$, which holds with probability at least $1-(WM)^{-\beta}$. Using (46), it becomes clear that the right hand side can be controlled with high probability by selecting $\Omega \geq C\beta\kappa\mu_{3,k-1}^2 R\max(W/M,1)\log(WM)$.

6.3 Proof of Lemma 4

Coherence of iterates $\boldsymbol{W}_k$ was defined earlier in (47). This Lemma is dedicated to showing that the coherence of the iterates is bounded. We start with matrix $\boldsymbol{W}_k$ and its coherences $\mu_{1,k}^2$ and $\mu_{2,k}^2$. The iterates are related as

$$\boldsymbol{W}_k = (\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathcal{P}_T)(\boldsymbol{W}_{k-1}).$$

Since $\boldsymbol{W}_k \in T$, we start with writing out

$$\begin{aligned}\boldsymbol{W}_k &= \kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k(\boldsymbol{W}_{k-1}) - \boldsymbol{W}_{k-1}\\ &= \sum_{(i,j)\in\Gamma_k}\kappa\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle\mathcal{P}_T(\boldsymbol{A}_{ij}) - \boldsymbol{W}_{k-1}\end{aligned}$$

The coherence $\mu_{3,k}^2$ of $\boldsymbol{W}_k$ is then

$$\mu_{3,k}^2 = \frac{M\Omega}{R}\max_{\substack{1\leq m\leq M\\ 1\leq\omega\leq\Omega}}\sum_{n\sim\mathcal{B}_\omega}\left(\sum_{(i,j)\in\Gamma_k}\kappa\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle[\mathcal{P}_T(\boldsymbol{A}_{ij})]_{(m,n)} - [\boldsymbol{W}_{k-1}]_{(m,n)}\right)^2 \tag{62}$$

We sart with bounding the following quantity with high probability

$$\eta_{mn}(\Gamma_k) := \left|\sum_{(i,j)\in\Gamma_k}\kappa\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle[\mathcal{P}_T(\boldsymbol{A}_{ij})]_{(m,n)} - [\boldsymbol{W}_{k-1}]_{(m,n)}\right| \tag{63}$$

using the scalar Bernstein bound. Instead of bounding $\{\eta_{mn}(\Gamma_k) > \epsilon\}$, we will bound the event $\{\eta_{mn}(\Gamma'_k) > \epsilon\}$, where the set Γ'_k is selected using Bernoulli model, i.e., we will bound the following quantity

$$\eta_{mn}(\Gamma'_k) := \left| \sum_{(i,j)\in\Gamma_k} \kappa\delta_{(i,j)} \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle [\mathcal{P}_T(\boldsymbol{A}_{ij})]_{(m,n)} - [\boldsymbol{W}_{k-1}]_{(m,n)} \right|$$

Let

$$Z_{ij} = \kappa\delta_{(i,j)} \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle [\mathcal{P}_T(\boldsymbol{A}_{ij})]_{(m,n)} - [\boldsymbol{W}_{k-1}]_{(m,n)}. \tag{64}$$

For this purpose, we need to calculate the variance

$$\sum_{(i,j)\in\Gamma'_k} \operatorname{E} Z_{ij}Z^*_{ij} \le \sum_{(i,j)} \kappa^2 \operatorname{E} \delta_{(i,j)} \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 [\mathcal{P}_T(\boldsymbol{A}_{ij})]^2_{(m,n)}$$

$$[\mathcal{P}_T(\boldsymbol{A}_{ij})]_{(m,n)} = [\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}]_{(m,n)} + [\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]_{(m,n)} - [\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]_{(m,n)}.$$

It follows that

$$[\mathcal{P}_T(\boldsymbol{A}_{ij})]^2_{(m,n)} \le 3\left([\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}]^2_{(m,n)} + [\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]^2_{(m,n)} + [\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]^2_{(m,n)} \right). \tag{65}$$

Using Lemma 5, we now calculate the variance term by term. The first term in the variance is

$$\begin{aligned} \sum_{(i,j)} \operatorname{E}[\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}]^2_{(m,n)} \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 &\le 3\sum_{i=1}^{M} \langle \boldsymbol{u}_m, \boldsymbol{u}_i\rangle^2 \cdot \sum_{j=1}^{\Omega}\sum_{\gamma\sim\mathcal{B}_j} \langle \tilde{\boldsymbol{e}}_\gamma, \tilde{\boldsymbol{e}}_n\rangle^2 \cdot \max_{ij} \sum_{\gamma\sim\mathcal{B}_j} W^2_{k-1}[i,\gamma] \\ &\le 3\|\boldsymbol{u}_m\|_2^2 \cdot \mu^2_{3,k-1} \frac{R}{M\Omega}, \end{aligned} \tag{66}$$

where the first inequality follows by the application of Lemma 6. For the second inequality, we have used the the definition of coherence in (21), and the fact that

$$\sum_{j=1}^{\Omega}\sum_{\gamma\sim\mathcal{B}_j} \langle \tilde{\boldsymbol{e}}_\gamma, \tilde{\boldsymbol{e}}_n\rangle^2 = 1$$

for any given index (m,n). Using Lemma 5, and 6 again, we obtain

$$\begin{aligned} \sum_{(i,j)} \operatorname{E}[\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]^2_{(m,n)} \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 &\le 3\sum_{(i,j)} \left(\sum_{\gamma\sim\mathcal{B}_j} W^2_{k-1}[i,\gamma] \cdot \sum_{\gamma\sim\mathcal{B}_j} \langle \boldsymbol{v}_\gamma, \boldsymbol{v}_n\rangle^2 \right) \langle \boldsymbol{e}_m, \boldsymbol{e}_i\rangle^2 \\ &= 3\max_{ij} \sum_{\gamma\sim\mathcal{B}_j} W^2_{k-1}[i,\gamma] \cdot \sum_{j=1}^{\Omega}\sum_{\gamma\sim\mathcal{B}_j} \langle \boldsymbol{v}_\gamma, \boldsymbol{v}_n\rangle^2 \cdot \sum_{i=1}^{M} \langle \boldsymbol{e}_m, \boldsymbol{e}_i\rangle^2 \\ &\le 3\max_{ij} \sum_{\gamma\sim\mathcal{B}_j} W^2_{k-1}[i,\gamma] \|\boldsymbol{v}_n\|_2^2 \cdot \|\boldsymbol{e}_m\|_2^2 \le 3\mu^2_{3,k-1} \frac{R}{M\Omega} \|\boldsymbol{v}_n\|_2^2. \end{aligned} \tag{67}$$

Similarly, the last variance term is

$$\begin{aligned} &\sum_{(i,j)} \operatorname{E}[\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]^2_{(m,n)} \langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle^2 \le \\ &\quad \le 3\sum_{i=1}^{M} \langle \boldsymbol{u}_n, \boldsymbol{u}_i\rangle^2 \cdot \sum_{j=1}^{\Omega}\sum_{\gamma\sim\mathcal{B}_j} \langle \boldsymbol{v}_\gamma, \boldsymbol{v}_n\rangle^2 \cdot \max_{ij} \sum_{\gamma\sim\mathcal{B}_j} W^2_{k-1}[i,\gamma] \\ &\quad \le 3\|\boldsymbol{u}_m\|_2^2 \cdot \|\boldsymbol{v}_n\|_2^2 \cdot \mu^2_{3,k-1} \frac{R}{M\Omega}. \end{aligned} \tag{68}$$

Using the fact that $\|\boldsymbol{v}_n\|_2^2 \leq 1$, we see that (67) dominates (68). Putting (66), (67), and (68) together, we have

$$\sigma_Z^2 = \sum_{(i,j)} \operatorname{E} Z_{ij} Z_{ij}^* \leq C\kappa (\|\boldsymbol{u}_m\|_2^2 + \|\boldsymbol{v}_n\|_2^2)\mu_{3,k-1}^2 \frac{R}{M\Omega}. \tag{69}$$

Now, we need to calculate the Orlicz-1 norm $\max_{ij} \|Z_{ij}\|_{\psi_1}$. Using standard argumnets in probability theory, see [46], we can show that

$$\|[\boldsymbol{U}\boldsymbol{U}^* \boldsymbol{A}_{ij}]_{(m,n)}\|_{\psi_2} \leq C\langle \boldsymbol{u}_m, \boldsymbol{u}_i\rangle \left(\sum_{\gamma \sim \mathcal{B}_j} \langle \tilde{\boldsymbol{e}}_\gamma, \tilde{\boldsymbol{e}}_n\rangle^2 \right)^{1/2} \leq C\|\boldsymbol{u}_i\|_2 \|\boldsymbol{u}_m\|_2$$

where the first inequality follows from the fact that for a fixed (m,n)

$$\sum_{\gamma \sim \mathcal{B}_j} \langle \tilde{\boldsymbol{e}}_\gamma, \tilde{\boldsymbol{e}}_n\rangle^2 \leq 1,$$

and that $\langle \boldsymbol{u}_m, \boldsymbol{u}_i\rangle \leq \|\boldsymbol{u}_i\|_2 \|\boldsymbol{u}_m\|_2 \leq \|\boldsymbol{u}_m\|_2$, as $\|\boldsymbol{u}_i\|_2 \leq 1$. Also,

$$\|[\boldsymbol{A}_{ij} \boldsymbol{V}\boldsymbol{V}^*]_{(m,n)}\|_{\psi_2} \leq C\langle \boldsymbol{e}_m, \boldsymbol{e}_i\rangle \left(\sum_{\gamma \sim \mathcal{B}_j} \langle \boldsymbol{v}_\gamma, \boldsymbol{v}_n\rangle^2 \right)^{1/2} \leq C\|\boldsymbol{v}_n\|_2,$$

the second inequality follows from the identity

$$\sum_{j=1}^{\Omega} \sum_{\gamma \sim \mathcal{B}_j} \langle \boldsymbol{v}_\gamma, \boldsymbol{v}_n\rangle^2 = \|\boldsymbol{v}_n\|_2^2,$$

and that $\langle \boldsymbol{e}_m, \boldsymbol{e}_i\rangle \leq 1$ for a fixed (m,n). Similaraly, we can show that

$$\|[\boldsymbol{U}\boldsymbol{U}^* \boldsymbol{A}_{ij} \boldsymbol{V}\boldsymbol{V}^*]_{(m,n)}\|_{\psi_2} \leq C\|\boldsymbol{u}_m\|_2 \|\boldsymbol{v}_n\|_2.$$

In addition, as before, we have

$$\|\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle\|_{\psi_2} \leq C \left(\sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma] \right)^{1/2}.$$

Using Lemma 7, Equation (64), (65), we can compute $\|Z_{ij}\|_{\psi_1}$ as follows

$$\begin{aligned}
&\max_{ij} \|Z_{ij}\|_{\psi_1}^2 \\
&\leq C\kappa^2 \max_{ij} \left(\|[\delta_{(i,j)} \boldsymbol{U}\boldsymbol{U}^* \boldsymbol{A}_{ij}]_{(m,n)}\|_{\psi_2}^2 + \|[\delta_{(i,j)} \boldsymbol{A}_{ij} \boldsymbol{V}\boldsymbol{V}^*]_{(m,n)}\|_{\psi_2}^2 + \|[\delta_{(i,j)} \boldsymbol{U}\boldsymbol{U}^* \boldsymbol{A}_{ij} \boldsymbol{V}\boldsymbol{V}^*]_{(m,n)}\|_{\psi_2}^2 \right) \|\langle \boldsymbol{A}_{ij}, \boldsymbol{W}_{k-1}\rangle\|_{\psi_2}^2 \\
&\leq C\kappa \max_{ij} \left(\|\boldsymbol{u}_n\|_2^2 + \|\boldsymbol{v}_m\|_2^2 \right) \cdot \sum_{\gamma \sim \mathcal{B}_j} W_{k-1}^2[i,\gamma] \\
&\leq C\kappa \left(\mu_1^2 \frac{R}{M} + \mu_2^2 \frac{R}{W} \right) \mu_{3,k-1}^2 \frac{R}{M\Omega}.
\end{aligned}$$

Then the Orlicz-norm term in the Bernstein bound is

$$U_1 \log\left(\frac{M\Omega \cdot U_1^2}{\sigma_Z^2} \right) \leq C\kappa \frac{R(\mu_1^2(W/M) + \mu_2^2)}{\Omega} \mu_{3,k-1}^2 \frac{R}{M\Omega} \log(WM).$$

Clearly, the Orlicz-norm term dominates the variance term in the Bernstein bound. Select $t = \beta \log(WM)$ in the Bernstein bound, which implies that

$$|\eta_{mn}(\Gamma_k')|^2 \leq C\beta\kappa (\|\boldsymbol{u}_m\|_2^2 + \|\boldsymbol{v}_n\|_2^2)\mu_{3,k-1}^2 \frac{R}{M\Omega} \log^2(WM)$$

$$\leq C\beta\kappa\frac{R(\mu_1^2(W/M)+\mu_2^2)}{\Omega}\mu_{3,k-1}^2\frac{R}{M\Omega}\log^2(WM)$$

holds with probability at least $1-(WM)^{-\beta}$. The second inequality follows from the definitions in (19), and (20). Using the bound on $|\eta_{mn}(\Gamma_k')|^2$, we can find a bound on the coherence of the iterates in (62), which is

$$\mu_{3,k}^2 \leq C\kappa\sum_{n\sim\mathcal{B}_\omega}(\mu_1^2\frac{R}{M}+\mu_2^2\frac{R}{W})\mu_{3,k-1}^2\log^2(WM)$$

Select $\Omega \geq C\beta\kappa R(\mu_1^2(W/M)+\mu_2^2)\log^2(WM)$, we can arrange for a constant C such that

$$\mu_{3,k}^2 \leq \frac{1}{2}\mu_{3,k-1}^2 \tag{70}$$

holds with probability at least $1-(WM)^{-\beta}$.

7 Auxiliary Lemmas for Theorem 1

Follwing Lemma will be useful in carying out several calculations in the proofs of the above lemmas.

Lemma 5. *Take $\boldsymbol{A}_{ij}$ as defined in* (38)*, and suppose $\boldsymbol{U}: M\times R$, and $\boldsymbol{V}: W\times R$ are orthogonal matrices with $\{\boldsymbol{u}_i\}_{1\leq i\leq M}$, $\{\boldsymbol{v}_i\}_{1\leq i\leq W}$ as rows, respectively. Then*

$$[\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}]_{(m,n)}^2 = \langle\boldsymbol{u}_m,\boldsymbol{u}_i\rangle^2\left(\sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\langle\tilde{\boldsymbol{e}}_\gamma,\tilde{\boldsymbol{e}}_n\rangle\right)^2, \tag{71}$$

$$[\boldsymbol{V}\boldsymbol{V}^*\boldsymbol{A}_{ij}^*]_{(m,n)}^2 = \langle\boldsymbol{e}_m,\boldsymbol{e}_i\rangle^2\left(\sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\langle\boldsymbol{v}_\gamma,\boldsymbol{v}_n\rangle\right)^2, \tag{72}$$

and

$$[\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]_{(m,n)}^2 = \langle\boldsymbol{u}_m,\boldsymbol{u}_i\rangle^2\left(\sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\langle\boldsymbol{v}_\gamma,\boldsymbol{v}_n\rangle\right)^2. \tag{73}$$

Proof. We will use the definition of $\boldsymbol{A}_{ij}$ in (38). Begin with

$$\begin{aligned}[\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}]_{(m,n)} &= \langle\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij},\boldsymbol{e}_m\tilde{\boldsymbol{e}}_n^*\rangle = \boldsymbol{e}_m^*\boldsymbol{U}\boldsymbol{U}^*\sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\boldsymbol{e}_i\tilde{\boldsymbol{e}}_\gamma^*\tilde{\boldsymbol{e}}_n\\ &= \sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\boldsymbol{e}_m^*\boldsymbol{U}\boldsymbol{u}_i\tilde{\boldsymbol{e}}_\gamma^*\tilde{\boldsymbol{e}}_n = \sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\boldsymbol{e}_m^*\boldsymbol{U}\boldsymbol{u}_i\tilde{\boldsymbol{e}}_\gamma^*\tilde{\boldsymbol{e}}_n\\ &= \langle\boldsymbol{u}_m,\boldsymbol{u}_i\rangle\sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\tilde{\boldsymbol{e}}_\gamma^*\tilde{\boldsymbol{e}}_n,\end{aligned}$$

Second,

$$\begin{aligned}[\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]_{(m,n)} &= \langle\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*,\boldsymbol{e}_m\tilde{\boldsymbol{e}}_n^*\rangle = \sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\boldsymbol{e}_m^*\boldsymbol{e}_i\tilde{\boldsymbol{e}}_\gamma^*\boldsymbol{V}\boldsymbol{V}^*\tilde{\boldsymbol{e}}_n\\ &= \langle\boldsymbol{e}_m,\boldsymbol{e}_i\rangle\sum_{\gamma\sim\mathcal{B}_j}d_i[\gamma]\langle\boldsymbol{v}_\gamma,\boldsymbol{v}_n\rangle.\end{aligned}$$

Finally,

$$[\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*]_{(m,n)} = \langle\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{A}_{ij}\boldsymbol{V}\boldsymbol{V}^*,\boldsymbol{e}_m\tilde{\boldsymbol{e}}_n^*\rangle$$

$$
\begin{aligned}
&= \boldsymbol{e}_m^* \boldsymbol{U}\boldsymbol{U}^* \left(\sum_{\gamma \sim \mathcal{B}_j} d_i[\gamma] \boldsymbol{e}_i \tilde{\boldsymbol{e}}_\gamma^* \right) \boldsymbol{V}\boldsymbol{V}^* \tilde{\boldsymbol{e}}_n \\
&= \sum_{\gamma \sim \mathcal{B}_j} d_i[\gamma] \boldsymbol{e}_m^* \boldsymbol{U} \boldsymbol{u}_i \boldsymbol{v}_\gamma^* \boldsymbol{V}^* \tilde{\boldsymbol{e}}_n \\
&= \langle \boldsymbol{u}_m, \boldsymbol{u}_i \rangle \left(\sum_{\gamma \sim \mathcal{B}_j} d_i[\gamma] \langle \boldsymbol{v}_\gamma, \boldsymbol{v}_n \rangle \right),
\end{aligned}
$$

□

Lemma 6. *Take $\boldsymbol{A}_{ij}$ as defined in* (38)*, and let $\boldsymbol{x}_k$, and $\boldsymbol{y}_k$ denote scalars or vectors. Then*

$$
\begin{aligned}
&\mathrm{E} \left(\sum_{\gamma,\gamma' \sim \mathcal{B}_j} d_i[\gamma] d_i[\gamma'] \boldsymbol{x}_\gamma \boldsymbol{x}_{\gamma'}^* \right) \left(\sum_{\gamma,\gamma' \sim \mathcal{B}_j} d_i[\gamma] d_i[\gamma'] \boldsymbol{y}_\gamma \boldsymbol{y}_{\gamma'}^* \right) = \\
&= \left(\sum_{\gamma \sim \mathcal{B}_j} \boldsymbol{x}_k \boldsymbol{x}_k^* \right) \left(\sum_{\gamma \sim \mathcal{B}_j} \boldsymbol{y}_\gamma \boldsymbol{y}_\gamma^* \right) + 2 \sum_{\gamma,\gamma'} \boldsymbol{x}_\gamma \boldsymbol{x}_{\gamma'}^* \boldsymbol{y}_\gamma \boldsymbol{y}_{\gamma'}^*.
\end{aligned}
$$

Proof. The proof of this Lemma is simple involves expanding and moving the expectation inside. We will use the result of this Lemma in cases when $\boldsymbol{x}_\gamma$, $\boldsymbol{y}_\gamma$ are both scalars and when one of these is a vector and other is a scalar. Furthermore, when both $\boldsymbol{x}_\gamma$, and $\boldsymbol{y}_\gamma$ are scalars then the result can be simplified using Cauchy-Schwartz inequality to yield

$$
\mathrm{E} \left(\sum_{\gamma,\gamma' \sim \mathcal{B}_j} d_i[\gamma] d_i[\gamma'] \boldsymbol{x}_\gamma \boldsymbol{x}_{\gamma'}^* \right) \left(\sum_{\gamma,\gamma' \sim \mathcal{B}_j} d_i[\gamma] d_i[\gamma'] \boldsymbol{y}_\gamma \boldsymbol{y}_{\gamma'}^* \right) \leq 3 \left(\sum_\gamma \boldsymbol{x}_\gamma^2 \right) \left(\sum_\gamma \boldsymbol{y}_\gamma^2 \right).
$$

□

Lemma 7. *Let X_1, and X_2 be two subgaussian random variables, i.e., $\|X_1\|_{\psi_2} < \infty$, and $\|X_2\|_{\psi_2} < \infty$. Then the product $X_1 X_2$ is a sub exponential random variable with*

$$
\|X_1 X_2\|_{\psi_1} \leq c \|X_1\|_{\psi_2} \|X_2\|_{\psi_2}.
$$

Proof. For a subgaussian random variable, the tail behaviour is

$$
\mathrm{P}\left\{|X| > t\right\} \leq e \exp\left(\frac{-ct^2}{\|X\|_{\psi_2}^2} \right) \qquad \forall t > 0;
$$

see, for example, [46]. We are interested in

$$
\begin{aligned}
\mathrm{P}\left\{|X_1 X_2| > \lambda\right\} &\leq \mathrm{P}\left\{|X_1| > t\right\} + \mathrm{P}\left\{|X_2| > \lambda/t\right\} \\
&\leq e \cdot \exp\left(-ct^2/\|X_1\|_{\psi_2}^2\right) + e \cdot \exp\left(-c\lambda^2/t^2 \|X_2\|_{\psi_2}^2\right).
\end{aligned}
$$

Select $t^2 = \lambda \|X_1\|_{\psi_2} / \|X_2\|_{\psi_2}$, which gives

$$
\mathrm{P}\left\{|X_1 X_2| > \lambda\right\} \leq 2e \cdot \exp\left(-c\lambda / \|X_1\|_{\psi_2} \|X_2\|_{\psi_2}\right).
$$

Now Lemma 2.2.1 in [47] imples that a random variable Z which obeys $\mathrm{P}\left\{|Z| > u\right\} \leq \alpha e^{-\beta u}$. Then $\|Z\|_{\psi_1} \leq (1+\alpha)/\beta$. Using this result, we obtain

$$
\|X_1 X_2\|_{\psi_1} \leq c \|X_1\|_{\psi_2} \|X_2\|_{\psi_2},
$$

which proves the result. □

8 Proof of Theorem 2

In this section, we show that the nuclear norm penalized estimators give ideal performance for the stable recovery of $\boldsymbol{X}_0$ in the presence of additive measurement noise. We will consider the measurement model in (12), and the noise characteristics in (25).

Proof. As will be clear later, the proof involves bounding the spectral norm:

$$\|\boldsymbol{\Theta}\| = \|\mathcal{A}^*(\boldsymbol{y}) - \operatorname{E}\mathcal{A}^*(\boldsymbol{y})\| \leq \|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_0)\| + \|\mathcal{A}^*(\boldsymbol{\xi})\| \tag{74}$$

The bound on $\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_0)\|$ can be obtained directly using Corollary 3, and the quantity $\|\mathcal{A}^*(\boldsymbol{\xi})\|$ is controlled using Lemma 8. Given these two results, the proof of Theorem 2 follows from the following main result in [41].

Theorem 4 (Oracle inequlaity in [41])**.** *Let $\hat{\boldsymbol{X}}$ be the solution of the* (24)*, and $\boldsymbol{X}_0 \in \mathbb{R}^{M\times W}$ matrix of rank R. If $\lambda \geq 2\|\boldsymbol{\Theta}\|$, then*

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_F^2 \leq \min\left\{2\lambda\|\boldsymbol{X}_0\|_*, 1.46\lambda^2 R\right\}.$$

Corollary 3. *Suppose Ω entries are observed using modulated multiplexing setup and let $\boldsymbol{Z}$ be a fixed $W \times M$ matrix with coherence μ_3^2 as defined* (21)*. Then for all $\beta > 0$,*

$$\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{Z})\| \leq C\|\boldsymbol{Z}\|_{\mathrm{F}}\sqrt{\frac{\mu_3^2\max(W/M,1)}{\Omega}}\sqrt{\beta\log(WM)}$$

with probability at least $1-(WM)^{-\beta}$ provided $\Omega \geq C\beta\log^2(WM)$.

Proof. The proof of the Corollry follows from Lemma 3 by selecting the number of partitions $\kappa = 1$. The result in the statement of the corollary is obtained from the bound in (61). In particular, the first term in the maximum in (61) dominates, when we select $\Omega \geq C\beta\log^2(WM)$. This proves the above corollary. □

Lemma 8. *Let $\boldsymbol{A}_{ij}$ be independent as defined* (38) *and pairs $(\boldsymbol{A}_{ij}, y_{ij})$ be independent. For $\beta > 1$ and $\Omega \geq C\min(W/M,1)\beta\log^2(WM)$, the following*

$$\|\mathcal{A}^*(\boldsymbol{\xi})\| \leq C\|\boldsymbol{\xi}\|_{\psi_2}\sqrt{\frac{\max(W/M,1)}{\Omega}}\sqrt{\beta\log(WM)},$$

holds with probability at least $1-(WM)^{-\beta}$ for a fixed constant C.

Using Lemma 2, and Lemma 8, we can bound (74), and obtain

$$\lambda \geq \sqrt{\frac{C\beta\{\|\boldsymbol{X}_0\|_{\mathrm{F}}^2\mu_3^2\max(W/M,1) + \|\boldsymbol{\xi}\|_{\psi_2}^2\max(W/M,1)\}\log(WM)}{\Omega}}$$

with probability at least $1-O(WM)^{-\beta}$ for a fixed constant C. Taking $\|\boldsymbol{X}_0\|_{\mathrm{F}} = 1$ without loss of generality, and $\Omega \geq C\beta\mu_3^2\max(W/M,1)\log^2(WM)$, which is in agreement with the assumptions on Ω in Corollary 3, and Lemma 8, we can control the right hand side. This proves Theorem 2. □

8.1 Proof of Lemma 8

The proof of this Lemma requires the use of matrix Bernstein's inequality 2. As it is required to bound the spectral norm of the sum $\mathcal{A}^*(\boldsymbol{\xi}) = \sum_{i,j}\xi_{ij}\boldsymbol{A}_{ij}$, we start with the summands $\boldsymbol{Z}_{ij} = \xi_{ij}\boldsymbol{A}_{ij}$. Because variables ξ_{ij} are zero mean, it follows that $E\boldsymbol{Z}_{ij} = \boldsymbol{0}$. The first quantity required is

$$\left\|\sum_{i,j}\operatorname{E}\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^*\right\| = \left\|\sum_{i,j}\operatorname{E}\xi_{ij}^2\cdot\operatorname{E}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^*\right\| \leq \max_{\substack{1\leq i\leq M\\1\leq j\leq\Omega}}\operatorname{E}\xi_{ij}^2\left\|\operatorname{E}\sum_{i,j}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^*\right\|.$$

Using the definition of $\boldsymbol{A}_{ij}$ in (38), we have

$$\begin{aligned}\left\|\sum_{i,j}\mathrm{E}\,\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^*\right\| &\leq \max_{\substack{1\leq i\leq M\\1\leq j\leq\Omega}}\mathrm{E}\,\xi_{ij}^2\left\|\sum_{i,j}\mathrm{E}\sum_{k,k'\sim\mathcal{B}_j}d_i[k]d_i[k']\boldsymbol{e}_i\tilde{\boldsymbol{e}}_k^*\tilde{\boldsymbol{e}}_{k'}\boldsymbol{e}_i^*\right\|\\ &= \max_{\substack{1\leq i\leq M\\1\leq j\leq\Omega}}\mathrm{E}\,\xi_{ij}^2\left\|\sum_{j=1}^{\Omega}\sum_{k\sim\mathcal{B}_j}\sum_{i=1}^{M}\boldsymbol{e}_i\boldsymbol{e}_i^*\right\|\\ &\leq W\max_{ij}\mathrm{E}\,\xi_{ij}^2 = \frac{(W/M)}{\Omega}\|\boldsymbol{\xi}\|_{\psi_2}^2, \end{aligned}\tag{75}$$

where the first equality follows from the independence of ξ_{ij} and $\boldsymbol{A}_{ij}$ and the last equality follows from (25). The second quantity required to calculate the variance (50) is

$$\begin{aligned}\left\|\sum_{i,j}\mathrm{E}\,\boldsymbol{Z}_{ij}^*\boldsymbol{Z}_{ij}\right\| &= \left\|\sum_{i,j}\mathrm{E}\,\xi_{ij}^2\cdot\mathrm{E}\,\boldsymbol{A}_{ij}^*\boldsymbol{A}_{ij}\right\|\\ &\leq \max_{ij}\mathrm{E}\,\xi_{ij}^2\left\|\sum_{i,j}\mathrm{E}\,\boldsymbol{A}_{ij}^*\boldsymbol{A}_{ij}\right\|\\ &= \frac{\|\boldsymbol{\xi}\|_{\psi_2}^2}{M\Omega}\cdot\left\|\sum_{i,j}\mathrm{E}\sum_{k,k'\sim\mathcal{B}_j}d_i[k]d_i[k']\tilde{\boldsymbol{e}}_k\boldsymbol{e}_i^*\boldsymbol{e}_i\tilde{\boldsymbol{e}}_{k'}^*\right\|\\ &= \frac{\|\boldsymbol{\xi}\|_{\psi_2}^2}{M\Omega}\left\|\sum_{i=1}^{M}\sum_{j=1}^{\Omega}\sum_{k\sim\mathcal{B}_j}\tilde{\boldsymbol{e}}_k\tilde{\boldsymbol{e}}_k^*\right\|\\ &= \frac{\|\boldsymbol{\xi}\|_{\psi_2}^2}{\Omega}\end{aligned}\tag{76}$$

Combining (75) and (76) and using (50) gives

$$\sigma_Z = \|\boldsymbol{\xi}\|_{\psi_2}\sqrt{\frac{\max(W/M,1)}{\Omega}}.$$

The final quantity required is the Orlicz norm of the summand matrices $\boldsymbol{Z}_{ij}$, i.e.,

$$\begin{aligned}\|\boldsymbol{Z}_{ij}\|_{\psi_2}^2 &= \|\xi_{ij}\|_{\psi_2}^2\|\boldsymbol{A}_{ij}\|^2\\ &\leq C\frac{\|\boldsymbol{\xi}\|_{\psi_2}^2}{M\Omega}\cdot\frac{W}{\Omega},\end{aligned}$$

then

$$\|\boldsymbol{Z}_{ij}\|_{\psi_2}\log^{1/2}\left(\frac{M\Omega\cdot\|\boldsymbol{Z}_{ij}\|_{\psi_2}^2}{\sigma_Z^2}\right)\leq C\sqrt{\|\boldsymbol{\xi}\|_{\psi_2}^2\frac{W/M}{\Omega^2}}\log^{1/2}(WM).$$

Hence, we obtain using $P=\Omega M$, and $t=\beta\log(WM)$ in the Bernstein's bound (51)

$$\left\|\sum_{ij}\xi_{ij}\boldsymbol{A}_{ij}\right\|\leq\max\left\{\|\boldsymbol{\xi}\|_{\psi_2}\sqrt{\frac{\max(W/M,1)}{\Omega}}\sqrt{\beta\log(WM)},\|\boldsymbol{\xi}\|_{\psi_2}\sqrt{\frac{W/M}{\Omega^2}}(\beta\log^{3/2}(WM))\right\}.$$

Select $\Omega\geq C\beta\min(W/M,1)\log^2(WM)$, then the first term dominates and the claim in Lemma 8 follows.

References

[1] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, March 2002.

[2] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[3] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[4] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[5] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Sig. Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[6] J. J. Fuchs, "Multipath time-delay detection and estimation," *IEEE Trans. Sig. Process.*, vol. 47, pp. 237–243, 1999.

[7] ——, "On the application of the global matched filter to doa estimation with uniform circular arrays," *IEEE Trans. Signal Process.*, vol. 49, no. 4, pp. 702–709, April 2001.

[8] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[9] S. Kunis and H. Rauhut, "Random sampling of sparse trigonometric polynomials," *Appl. Comp. Harmon. Analysis*, vol. 22, 2006.

[10] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Comm. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, 2008.

[11] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Appl. Comp. Harm. Analysis*, vol. 35, no. 1, pp. 111–129, July 2013.

[12] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inform. Theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[13] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Comm. Pure Appl. Math.*, vol. 67, no. 6, pp. 906–956, June 2014.

[14] Ali Ahmed and Justin Romberg, "Compressive multiplexing of correlated signals," *IEEE Trans. Inform. Theory*, vol. 1, pp. 479–498, 2015.

[15] A. Hormati, O. Roy, Y. M. Lu, and M. Vetterli, "Distributed sampling of signals linked by sparse filtering: Theory and applications," *IEEE Trans. Sig. Process.*, vol. 58, no. 3, pp. 1095–1109, 2010.

[16] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," *arXiv preprint arXiv:0901.3403*, 2009.

[17] M. Mishali and Y. Eldar and O. Dounaevsky and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits Devices Syst.*, vol. 5, no. 1, pp. 8–20, 2011.

[18] M. Mishali and Y. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Sig. Process.*, vol. 57, no. 3, pp. 993–1009, 2009.

[19] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Trans. Sig. Process.*, vol. 59, no. 10, pp. 4719–4734, 2011.

[20] W. Mantzel and J. Romberg, "Compressed subspace matching on the continuum," *arXiv preprint arXiv:1407.5234*, 2014.

[21] H. Malvar and D. Staelin, "The LOT: Transform coding without blocking effects," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 553–559, April 1989.

[22] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[23] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[24] D. Slepian, "On bandwidth," *Proceedings of the IEEE*, vol. 64, no. 3, pp. 292–300, March 1976.

[25] ——, "Prolate spheroidal wave functions, fourier analysis and uncertainty v — the discete case," *Bell Systems Tech. Journal*, vol. 57, pp. 1371–1430, 1978.

[26] Schlottmann, C.R. and Hasler, P.E., "A highly dense, low power, programmable analog vector-matrix multiplier: The fpaa implementation," *IEEE J. Emerg. Sel. Topic Circuits Sys.*, vol. 1, no. 3, pp. 403–411, 2011.

[27] Chawla, R. and Bandyopadhyay, A. and Srinivasan, V. and Hasler, P., "A 531 nw/mhz, 128× 32 current-mode programmable analog vector-matrix multiplier with over two decades of linearity," in *Proc. IEEE Conf. Custom Integr. Circuits.*, 2004, pp. 651–654.

[28] J. Tropp and J. Laska and M. Duarte and J. Romberg, and R. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 520–544, 2010.

[29] J. Laska and S. Kirilos and M. Duarte and T. Raghed and R. Baraniuk and Y. Massoud, "Theory and implementation of an analog-to-information converter using random demodulation," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2007, pp. 1959–1962.

[30] J. Yoo, S. Becker, M. Loh, M. Monge, and E. Candès, "A 100MHz-2GHz 12.5x sub-Nyquist rate receiver in 90nm CMOS," in *Proc. IEEE Radio Freq. Integr. Circuits Symp. (RFIC)*, 2012.

[31] J. Yoo, C. Turnes, E. Nakamura, C. Le, S. Becker, E. Sovero, M. Wakin, M. Grant, J. Romberg, A. Emami-Neyestanak, and E. Candès, "A compressed sensing parameter extraction platform for radar pulse signal acquisition," *Submitted to IEEE J. Emerg. Sel. Topics Circuits Syst.*, February 2012.

[32] T. Murray, P. Pouliquen, A. Andreou, and K. Lauritzen, "Design of a CMOS A2I data converter: Theory, architecture and implementation," in *Proc. IEEE Annu. Conf. Inform. Sci. Syst. (CISS)*, Baltimore, MD, 2011, pp. 1–6.

[33] J. Romberg, "Compressive sensing by random convolution," *SIAM J. Imag. Sci.*, vol. 2, no. 4, pp. 1098–1128, 2009.

[34] J. Haupt and W. Bajwa and G. Raz and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5862–5875, 2010.

[35] H. Rauhut and J. Romberg and J. Tropp, "Restricted isometries for partial random circulant matrices," *Appl. Comput. Harmonic Anal.*, vol. 32, no. 2, pp. 242–254, 2012.

[36] J. Tropp and M. Wakin and M. Duarte and D. Baron and R. Baraniuk, "Random filters for compressive sampling and reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, 2006.

[37] E. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[38] K. Mohan and M. Fazel, "New restricted isometry results for noisy low-rank recovery," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Austin, Texas, June 2010.

[39] B. Recht, “A simpler approach to matrix completion,” *J. Mach. Learn. Res.*, vol. 12, pp. 3413–3430, 2011.

[40] A. Ahmed and B. Recht and J. Romberg, “Blind deconvolution using convex programming,” *IEEE Trans. Inform. Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[41] V. Koltchinskii, K. Lounici, and A. Tsybakov, “Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion,” *Ann. Stat.*, vol. 39, no. 5, pp. 2302–2329, 2011.

[42] M. Fazel and E. Candès and B. Recht and P. Parrilo, “Compressed sensing and robust recovery of low rank matrices,” in *Proc. 42nd IEEE Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, 2008, pp. 1043–1047.

[43] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” *Ann. Stat.*, pp. 1302–1338, 2000.

[44] M. Ledoux, *The concentration of measure phenomenon.* AMS, 2001, vol. 89.

[45] J. Tropp, “User-friendly tail bounds for sums of random matrices,” *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[46] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *Compressed Sensing: Theory and Applications*, pp. 210–268, November 2012.

[47] A. V. der Vaart and J. Wellner, *Weak Convergence and Empirical Processes.* Springer, 1996.